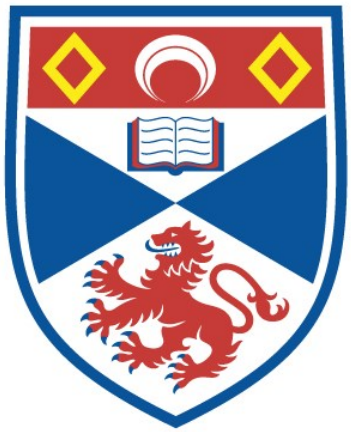
SEX DIFFERENCES IN SOCIAL LEARNING: EXPLORING THE LINKS WITH RISK AVERSION AND CONFIDENCE

Charlotte Olivia Brand

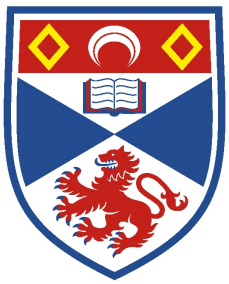
A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2017

Sex Differences in Social Learning: Exploring the links with Risk Aversion and Confidence

Charlotte Olivia Brand



University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of
PhD
at the
University of St Andrews

December
2016

1. Candidate's declarations:

I, Charlotte Brand, hereby certify that this thesis, which is approximately 40,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in September, 2013 and as a candidate for the degree of Ph.D. in December 2016; the higher study for which this is a record was carried out in the University of St Andrews between 2013 and 2016.

Date: 20/12/16

Signature of candidate:

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Ph.D. in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date: 20/12/16

Signature of supervisor:

3. Permission for publication:

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

PRINTED COPY: No embargo on print copy

ELECTRONIC COPY: No embargo on electronic copy

Date: 20/12/16

Signature of candidate:

Signature of supervisor:

Contributions:

Chapter 2:

Ken Munro from the School of Psychology & Neuroscience, University of St Andrews coded the experimental spaceship game according to my design and instructions. I modified the code to tweak aspects of the experiment once it had been initially coded. The concept, experimental design, running of the experiment, data collection, data analysis and interpretation was all conducted by myself.

Chapter 3:

James Ounsley and Dr. Daniel van der Post from the School of Biology, University of St Andrews helped to write the initial analysis code in Jags for Study 1. All subsequent code was written by me. The stimuli for the experiment were obtained from Gerianne Alexander, from Texas A&M University, on request. I modified the stimuli, designed the experiment, ran the experiments, collected data and analysed data myself.

Chapter 4:

Dr. Thomas Morgan from University of California, Berkely, wrote the simulation code for the stereotype threat model. The idea was conceived of by myself and James Ounsley. I ran the simulations, analysed the data, plotted the results and interpreted the results myself.

Funding:

My work was funded by the John Templeton Foundation Grant: 'Exploring the Evolutionary Origins of Cultural Complexity, Creativity and Trust' awarded to Prof. Andrew Whiten and Prof. Kevin Laland, from October 2013 - September 2016.

Statistical training:

I acquired statistical training during this time by attending a course on "Applied Bayesian Modelling for Ecologists and Epidemiologists" 26th - 31st October 2015, operated by *PR Statistics.* I supplemented this training by following the *Statistical Rethinking* course material by Richard McElreath at http://xcelab.net/rm/statistical-rethinking/

Acknowledgements:

Thank you to my supervisors, Gillian Brown and Kate Cross, who both went above and beyond the role of a supervisor by being so caring, supportive and understanding throughout my PhD, withstanding many bursts of tears and frustration along the way. Thanks to Gillian, who encouraged me to apply for a PhD in the first place, and who supported me in spending a little too much time reading about statistics when I could have been running more experiments! Thanks to Kate who not only provided me with infinite amounts of tea and giggles, but also inspired me to be a better communicator, feminist and scientist. I am so grateful to both for encouraging and helping me in my postdoc applications, and doing all they could to help me rapidly write-up my thesis in time to start my postdoc at the other end of the country!

Thank you to all of the junior members of the JTF funded project (aka JTF Crew): Lewis Dean, Amanda Lucas, Emily Burdett, Gill Vale, Jennifer Botting, Rachel Harrison, Stuart Watson, Sarah Davis, Keelin Murray, Joseph Stubbersfield, Wataru Toyokawa, Christopher Ritter, Ana Navarrete, Daniel van der Post (DvdP), Daniel Cownden (CanaDan), James Ounsley, Murillo Pagnotta, Elena Miu, Andrew Whalen, Camille Troisi, and their predecessors: Thomas Morgan, Sally Street, Cara Evans and Alice Cowie. You were such a wonderful, inspiring bunch to do science with, to drink merrily with, and to get to know over the years. To Tom and Sally in particular, for dazzling me with your R skills and Bayesian knowledge in my first few weeks as a PhD student, motivating me to try and follow in your ginormous footsteps!

A special mention has to go to James Ounsley, who was not only my housemate and landlord, but also my peer, collaborator, statistics tutor, chef, music advisor, life coach and best friend. My PhD may have turned out very differently without the patient guidance and support of James, who spent a few evenings too many staying up until 1am explaining how the MCMC algorithm works just one more time… or going through the past ten versions of a jags file to find a bug… or debating the meaning of maths, the universe and everything. Thank you for your patience, wisdom, encouragement and belief in me, I owe you many beers.

A special thanks to Elena Miu, with whom I had the pleasure of enjoying lengthy, in depth discussions on everything from the definition of conformity, to Kevin's theories of language evolution, from friendship and family, to love and lust. We covered it all, and put the world to rights on almost a daily basis. I'll miss you terribly and I admire and respect you too much to say.

To Dugald Foster, who has reignited my passion for science and life since we met at EHBEA, 2016. He has filled the last few months of my PhD with more love, care, support, kindness, fun and frolicking than I thought was possible. My personal sandwich-maker, tea-brewer, pot-washer, alarm clock, chef, reference manager, thesaurus, tear-wiper and much more. Thank you, Dougie.

Finally to my parents, who, despite still not quite understanding what a PhD is or why I am doing one, have always fully supported and encouraged me to do whatever intrigues and captivates me. Never questioning my motivations or choices and always proud of every tiny achievement, I know it's thanks to them that I have been able to pursue an academic career studying human evolution- something which still fills me with excitement and satisfaction on a daily basis.

Abstract

In this thesis, I explore sex differences in adult human social learning, and how these sex differences might be underpinned by differences in risk-taking and confidence. The capacity for high-fidelity social learning is fundamental to the complex culture observed in humans. Examining when we choose to learn socially rather than asocially and the factors that influence these choices is valuable for understanding broader patterns of cultural evolution. In **Chapter 1,** I give an overview of the literature on sex differences in social learning, particularly how individual differences in risk-taking and confidence impact upon social information use. I introduce stereotype threat as a possible factor affecting women's confidence. I then use the stereotype threat literature as an example of the replication crisis in psychology, and discuss how improved methods of statistical analysis could help to elucidate the ambiguity in this literature. **Chapter 2** provides an experimental investigation into when adult humans choose to use social or asocial information to solve a virtual construction task. I found that when the asocial information was made to be risky, women, but not men, preferred to learn socially. In **Chapter 3,** I present experimental data exploring the links between confidence and conformity. I found that lower confidence led to a greater likelihood of conforming, although I was unable to reliably alter confidence levels experimentally. In **Chapter 4,** I present a model of the stereotype threat literature by simulating numerous datasets and analysing them using four different statistical methods. I found that using informative priors in a Bayesian framework provided greater certainty about the presence or absence of an effect in a population. Finally, in **Chapter 5** I conclude that studying the factors that lead to sex differences in social learning provides researchers with a greater understanding of the dynamics of cultural evolution.

Chapter 1: Introduction 9
1.1 Overview 10
1.2 Terminology and Methodology 13
1.3 Sex Differences in Social Learning 19
1.4 Conformity and Confidence 31
1.5 Uncertainty in the Stereotype Threat Effect 41
Thesis Outline 47
Chapter 2: 48
Risk-taking and Social Learning 48
2.1 Introduction 49
2.2 Methods 52
2.3 Analysis 62
2.4 Results 64
2.5 Discussion 71
Chapter 3: 76
Confidence and Conformity 76
Introduction 77
Study 1: Manipulating Confidence 81
3.1 Methods 81
3.1 Results 93
3.1 Discussion 98
Study 2: Confidence and Conformity 101
3.2 Methods 101
3.2 Results 105
3.2. Discussion 111
Study 3: Replication of Study 2 114
3.3 Method 114
3.3 Results 115
3.3 Discussion 122
General Discussion 123
Chapter 4: 127
Statistical Comparisons of the Stereotype Threat Literature 127
4.1 Introduction 128

4.2 Methods 130
4.3 Results 135
4.4 Discussion 147
Chapter 5: Discussion 152
5.1 Summary of findings 153
5.2 Implications of sex differences in social and asocial learning 155
5.3 Implications of sex differences in confidence and conformity 157
5.4 Implications of Bayesian analysis methods in psychology 160
5.5. Future work 162
5.6 Conclusion 164
References 166
Appendix 194

Chapter 1: Introduction

1.1 Overview

The field of cultural evolution is based on the assertion that human culture is governed by similar rules to those that underlie biological evolution, and thus can be understood and explained in terms of Darwinian evolutionary processes (Mesoudi, Whiten & Laland 2004). In this way, human culture can change over time as successful behaviours persist over generations and unsuccessful ones die out, analogously to biological evolution. Behaviours are modified and passed between individuals through social learning, in which individuals acquire new behaviours through observing or interacting with conspecifics (Hoppitt & Laland 2013). Humans are proficient social learners, and it has been argued that our capacity for high-fidelity social learning is one of the reasons that human culture is so vast, complex and diverse (Dean, Kendal, Schapiro, Thierry & Laland 2012). A large amount of research effort has gone into investigating the factors that lead to human culture far exceeding the complexity of animal cultures (Dean, Vale, Laland, Flynn & Kendal 2014). Other research questions are concerned with the cultural micro-evolutionary processes in humans specifically, i.e. the details of who, when, and how people learn from others (Mesoudi 2016). In particular, I am interested in whether and when sex differences emerge in social learning, and how these differences might be underpinned by individual differences in risk-taking and confidence.

There is some sparse evidence of sex differences in social learning from studies of non-human animals (e.g. Lonsdorf, Eberly & Pusey 2004) and from some human experimental work in children and adults (e.g. Flynn & Whiten 2008), and I am interested in how these differences might relate to other factors such as risk-taking and confidence. Theoretical and experimental work suggests that when personal experience becomes unreliable and risky, individuals rely on social information (e.g. Arbilly, Motro, Feldman & Lotum 2011; van Bergen, Coolen & Laland 2004). Additionally, individual differences in risk-taking tendencies predict the use of social or asocial learning to solve a problem (see

Brosnan & Hopper 2014). Thus, due to reported sex differences in men and women's risk-taking tendencies (Charness & Gneezy 2012), I am interested in whether men and women rely on risky sources of social and asocial information to differing extents. Similarly, differences in confidence have been found to underpin differences in conformity in previous studies, in that lower levels of confidence in a task led to higher levels of conformity (Morgan et al. 2012). Moreover, a sex difference in confidence on a mental rotation task had an indirect effect on conformity in the task (Cross et al. 2016). The factors that influence men and women's confidence in a task may themselves be socially learned or part of the cultural environment. One such example is stereotype threat (Stoet & Geary 2012), which can be seen as a type of social information affecting members of a stereotyped group in particular. Therefore I investigated whether manipulating men and women's confidence using stereotype threat differentially affected their likelihood of conforming to a majority decision.

Distinguishing when men and women use social and asocial learning to differing extents, and deciphering the contexts in which factors such as confidence and risk-taking play an important role, are valuable for understanding and predicting patterns of social information use. A fuller understanding of these cultural micro-evolutionary processes can inform research that looks at the macro-evolutionary consequences, such as the population-level dynamics of social learning (Mesoudi 2016). In addition, understanding how confidence and risk-taking differences affect men and women's social learning behaviour can also shed light on why men and women behave differently in other contexts, as well as highlighting the importance of exploring underlying factors that account for reported sex differences in behaviour. By manipulating the factors that differentially affect men and women's behaviour, sex differences in behaviour may be exaggerated, ameliorated or remain unchanged, giving greater insights into how flexible or context-dependent behavioural sex differences may or may not be.

In this thesis, I present experimental work investigating whether altering the riskiness of social and asocial information differentially affects men and women's tendency to use these information sources **(Chapter 2).** I also present an attempt to manipulate men and women's confidence using stereotype threat, to see if a decrease in confidence results in an increase in conformity (**Chapter 3).** I then present a study addressing the uncertainty within the stereotype threat literature by simulating datasets based on this literature and analysing them using a variety of statistical methods (**Chapter 4).** Before discussing this work in detail, I will first give a broad overview of the relevant literature. Specifically, I highlight the current experimental and empirical evidence for sex differences in social learning in animals and humans, and how these sex differences in social learning might relate to individual differences in risk-taking. I then discuss the conformity literature from a social psychology and cultural evolution perspective, as well as discussing the current evidence for sex differences in conformity. I point to research suggesting that sex differences in conformity might be underpinned by differences in confidence, and how confidence itself may be influenced by social information such as stereotypes. Finally, I describe the controversies surrounding the stereotype threat literature, the broader context of the replication crisis in psychology, and how these controversies might be addressed with improved statistical methods.

1.2 Terminology and Methodology

1.2.1 Sex and Gender

Throughout this thesis I have used both the term ‘sex’ and ‘gender’ when discussing my own and others’ research. The terms ‘sex’ and ‘gender’ can carry different meanings and interpretations depending on not only the historical context in which the term is being used, but also the norms and cultures that have developed within certain disciplines and fields of study (Torgrimson & Minson 2005). For example, the term ‘gender’ was not widely used in Psychology journals before the 1960s (Muehlenhard & Peterson 2011), and has markedly increased in frequency in both Physiology and Psychology journals since then (Torgrimson & Minson 2005; Muehlenhard & Peterson 2011).

The term ‘gender’ became more commonplace after John Money and colleagues (Money et al. 1955; 1957) defined *gender role* as distinct from anatomical or physiological sex (e.g. chromosomes, hormones, reproductive organs and external genitalia) and childhood socialisation (e.g. assigned sex and rearing) in the context of research into individuals with ambiguous biological sex, then termed *hermaphrodites,* now more commonly termed *intersexed.* From then on, many other researchers, notably Gayle Rubin (1975) and Rhoda Unger (1979) proposed a move away from using the term ‘sex’ ubiquitously and replacing it with the term ‘gender’ when appropriate. Specifically, Rubin argued that ‘sex’ refers to the biological body which one is born with, whereas ‘gender’ is the social role that is imposed on each sex (Rubin 1975). Further to this, Unger proposed that when researchers use the term ‘sex’ it is implied, or can be construed as meaning that there is an underlying biological mechanism responsible for any characteristics or differences specific to one sex (Unger 1979). She argued that using the term ‘gender’ instead would make it less likely for people to assume that psychological differences are the result of physiological sex differences, rather than environmental or social factors. Since Unger’s publication, the use of the term ‘gender’

has rapidly increased relative to the term ‘sex’ (Glasser & Smith 2008; Basow 2010; Muehlenhard & Peterson 2011). Indeed a recent article on ‘The New Science of Cognitive Sex Differences’ by Miller and Halpern (2014, p.37) explicitly defined ‘gender’ as “meanings that societies and individuals give to female and male categories”, whereas ‘sex’ was defined as “grouping of people into female and male categories. Use of the term ‘sex differences’ does not imply that those differences are primarily biologically or environmentally caused.”

Despite strong arguments for favouring one term over the other in certain contexts, in investigating current psychological researchers’ uses and definitions of the terms sex and gender, Muehlenhard and Peterson (2011) found many inconsistencies in not only the way in which terms were used, but also how they were defined, as well as justifications for those definitions. The authors concluded, similarly to Yoder (2003), that, although distinguishing between sex and gender has been important in developing understanding of the relative contributions of physiological versus societal factors in shaping one’s sex identity, the reality is that these lines cannot be distinctly drawn and are of course blurred. Thus, strict definitions distinguishing between the relative physiological and societal contributions to one’s sex are inevitably going to be inconsistent, incomplete or contradictory. Yoder argued that if we let go of the distinction between sex and gender, this will “open the door to regarding sex and gender as inseparable and intertwined so that a holistic understanding of women and men, girls and boys, will include biology (sex) and what our culture makes of our biological sex (i.e. gender)” (Yoder 2003, p.17).

In this thesis, I explicitly look to both evolutionary and social factors in helping to understand sex differences in human behaviour. In particular I employ ideas from social psychology, evolutionary psychology and cultural evolution which inherently reject a dichotomy between biology and culture, but fundamentally acknowledge their complex intertwinement. For example, in Chapter 2 I consider the idea that sex differences in risk-taking may affect social

learning. Although I refer to comparative literature showing sex differences in risk-taking to be present in non-human animals, I assume that human sex differences in risk-taking are determined by a multitude of environmental, physiological, evolutionary and cultural factors. Similarly in Chapter 3, I explore how gender stereotypes may affect men and women's confidence in mental rotation tasks. Again, I assume that sex differences in confidence are determined by a combination of environmental, physiological, evolutionary and cultural factors. Thus, in line with Yoder, I try to avoid the false "biology versus culture" dichotomy by assuming that both factors are always at play in the emergence of any behaviour. Therefore I use the terms 'sex' and 'gender' interchangeably throughout this thesis; instances in which 'gender' are used reflect the norm of the field of literature to which I am referring (e.g. social psychology) whereas 'sex' is used in particular when referring to comparative animal behaviour studies.

1.2.2 Online studies and external/internal validity

In this thesis I have used computer-based tasks to examine the effects of factors such as confidence and risk-taking on individuals' social learning propensities. Some of these tasks were conducted online via an online survey and recruiting service, Qualtrics. Other tasks were conducted by recruiting participants to take-part in a computer-based study in person.

A particular concern with recruiting participants online is one of internal validity. Due to the participants being remote from the experimenter, and indeed never meeting the experimenter, it is difficult to gain confidence in the participants' motivation, attention and commitment to the online task. For example, participants may be skim-reading the instructions and/or not thoroughly understanding the task, thus not behaving as they would if

they were participating with full understanding of the consequences of their actions in the task. However, it is important to note that all tasks designed and implemented using Qualtrics (all tasks in Chapter 3) include specific attention filters and measures to exclude participants who are not giving the task appropriate attention. For example, any participants who took less than a third of the average amount of time taken to complete the task are automatically excluded by Qualtrics' own quality filters, as this suggests that the participant was completing the task as quickly as possible rather than fully engaging with the task. My study also made use of time-limits per question, so participants who did not answer a question within the allotted time were excluded also (this prevents participants in engaging in other activities such as watching T.V. whilst taking part). Another measure that is put in place by Qualtrics is the use of attention filters within the task itself. These consist of up to three randomly placed requests in the task, such as the request to "write the following word in the box below" to ensure participants are reading the information thoroughly throughout. Thus, although researchers may never know for certain how online participants interact with their task, these measures can at least provide confidence in the attention that participants are giving to the task.

As the use of online platforms such as Mturk become more popular, researchers are testing the validity of these samples compared to other commonly used samples including face-to-face recruited samples. Indeed, growing research is beginning to suggest that Mturk samples actually pay more attention (Berinsky, Huber & Lens 2012; Hauser & Shwarz 2016); are more representative of the general population (Berinsky, Huber & Lens 2012; Poalacci & Chandler 2014; Huff & Tingley 2015) and respond similarly to external stimuli as other samples (Berinsky, Huber & Lens 2012). One of the main differences found between Mturk samples and the general public is that they tend to be younger and more ideologically liberal (Berinsky, Huber & Lens 2012; Poalacci & Chandler 2014). Accordingly, researchers are encouraged to bear these caveats in mind when interpreting results of Mturk studies. One

way to boost confidence in the results of online studies would be to also collect data with participants in person, to see if they respond comparably.

Recruiting participants in person can help to control for some of the problems discussed above, for example participants are able to ask the experimenter questions if they do not understand something, and experimenters can monitor and oversee participants' engagement and attention with the task. However, recruiting participants in person can have its own obstacles and restrictions. Notoriously, participants recruited by university researchers are often, due to accessibility, recruited from within the university, usually undergraduate students. This process can lead to a sample that is more homogenous and less varied than a typical sample of the general public. These homogenous samples have been aptly dubbed by the acronym 'WEIRD', which stands for western, educated, industrialised, rich and democratic (Henrich, Heine & Norenzayan 2010). This is a well-known and commonly encountered problem in the field of Psychology and Anthropology thus researchers are actively looking for alternative samples, often through online resources. Therefore, a good strategy for any behavioural research is to obtain a variety of samples with which to conduct a study, in order to test their replicability across samples.

Another concern with results from computer-based experiments is of external validity, i.e. do participants interaction with a computer task reflect how people would act in the real world, with real people. Furthermore, when playing an isolated computer game the consequences of participants' actions rarely affect anything other than their final score or payment. Therefore participants may not be motivated to behave in a way in which they would if there were 'real world' consequences to their actions. However, as the famous expression by George E. P. Box states "all models are wrong, some are useful." To understand complex, real world behaviour, scientists inevitably have to break down their study system into simpler

units to study. Thus the argument that online experiments may not translate into real life may be similarly applied to many other experimental set-ups in behavioural science, online or not. However, it is imperative to bear this caveat in mind, especially when interpreting results of a single study. Again, if many similar experiments are conducted on different aspects of a given behaviour and provide replicable results, it may be reasonable to assume that they are addressing a general behavioural mechanism that applies in many contexts, including ‘real world’ scenarios. Indeed, many behavioural researchers are starting to address the problem of external validity in their research studies by using observational data from local communities or social media in order to study more ‘real world’ scenarios.

1.3 Sex Differences in Social Learning

1.3.1 Darwinian Cultural Evolution

As previously mentioned, the field of cultural evolution is based on the assertion that human culture is governed by similar rules to those that underlie biological evolution, and thus can be understood and explained in terms of Darwinian evolutionary processes (Mesoudi, Whiten & Laland 2004). However, the extent to which cultural evolution can be described as truly Darwinian has been heavily debated. Culture can be characterized as changing over time under the same core principles of Darwinian natural selection, i.e. variation, competition and inheritance (Mesoudi, Whiten & Laland 2004). Nevertheless, cultural transmission can clearly differ from biological replication in many ways, most conspicuously in that culture can be transmitted horizontally or obliquely rather than just vertically (Mesoudi, Whiten & Laland 2006). Many other parallels and distinctions have been drawn as the field of cultural evolution has gained recognition and prominence (see Mesoudi 2007 for review). A related issue is whether the field of biological evolution needs to incorporate cultural evolution formally into its framework, namely as part of an “Extended Evolutionary Synthesis,” (Laland, Wray & Hoekstra 2014). The details of this debate are beyond the scope of this thesis, however it is worth noting that a growing group of evolutionary biologists argue that the process of biological evolution goes beyond those of the modern synthesis, and that we need to emphasise and better understand other processes including niche construction, cultural evolution and developmental bias (Laland, Wray & Hoekstra 2014).

Although much valuable work has been done in the past 30 years on the general population-level processes and consequences of cultural evolution, as well as understanding the scope and mechanisms of social learning in humans and other species, the field of cultural evolution has been criticized for neglecting cognitive science and the concept of 'meta-cognition' (Heyes 2016; Clarke & Heyes 2017; Heyes 2017). In particular, it has been argued that many of the social learning 'strategies' may be better explained by domain-general associative learning processes (Heyes 2012). Much work has been done in bringing together evolutionary biology, evolutionary anthropology, evolutionary psychology, comparative psychology and social psychology within the cultural evolution framework (Henrich & McElreath 2003; Mace & Holden 2005; Mesoudi, Whiten & Laland 2006; Mesoudi 2009; Mesoudi 2011; Mesoudi 2016). However the next step for the field of cultural evolution could be to combine theories and evidence from within cognitive neuroscience research to understand how the functioning of the human brain enables our species to be 'cultural learning' experts as outlined by many prominent cultural evolutionists (Henrich 2016; Laland 2017). Appealing to cognitive neuroscientists could also help shed light on the "cultural attractors" debate that is currently ongoing within cultural evolutionary research also (see Acerbi & Mesoudi 2015).

1.3.2 Social Learning Strategies

A core component of cultural evolutionary theory is the use of social learning by humans and non-human animals, as it is the mechanism by which cultural information is transmitted from one individual to another. Social learning has been defined as "learning that is facilitated by observation of, or interaction with, another individual or its products" (Hoppitt & Laland 2013, p.4). In contrast, *asocial* learning is broadly defined as "individual learning stemming solely

through direct interaction with the environment, for example, through trial and error" (Rendell et al. 2010, p.209). Thus, in classic 'producer scrounger' dilemmas (Barnard & Sibly 1981), asocial learners are seen as the information producers, and social learners as the information scroungers (e.g. Kameda & Nakanishi 2002). Social learners gain information about the environment that has been acquired by others, thus avoiding the costs of directly sampling from the environment themselves. A population cannot entirely consist of social learners, as some individuals have to gain new knowledge directly from the environment (Rogers 1988). Moreover, copying others indiscriminately will not inherently lead to fitness benefits; and social learning only becomes adaptive when it is used selectively and in combination with asocial learning (Laland 2004). Therefore, identifying when individuals choose to learn socially or asocially is a valuable pursuit.

A social learning strategies tournament attempted to tease apart when it is adaptive to copy, and who to copy from (Rendell et al 2010). The results of this tournament suggested that strategies relying heavily on social learning are the most successful, even when asocial learning is no more costly (Rendell et al. 2010). Although theoretical models predict when it may be most adaptive to use social and asocial information, other factors may come into play that influence when individuals employ social or asocial learning. Some experimental studies with adult participants have found that social learning is often under-used, and asocial learning is favoured even when social learning is the adaptive choice (Mesoudi 2011; Atkinson, O'Brien & Mesoudi 2012; Morgan et al. 2012). The propensity to choose asocial over social information has also been associated with cultural differences in individualism and collectivism, in that Western cultures are typically considered to be more individualistic than East Asian cultures, which are relatively collectivist in comparison (Mesoudi et al. 2015). Indeed differences in prosociality, narcissism, individualism and collectivism have been found to account for individual differences in the tendency to use asocial as opposed to

social information (Toelch et al. 2014; Eriksson & Strimling 2009). An additional factor is risk; if either social or asocial learning becomes a more risky strategy than the other, this difference in riskiness could influence an individual's likelihood to use social or asocial learning. Similarly, individual differences in risk-taking may determine which individuals use social or asocial learning in contexts when the pay-offs differ between the options. How differences in risk-taking interact with social and asocial learning is discussed further below.

1.3.3 Risk-taking and Asocial Learning

Asocial learning relies on performing a novel behaviour that has not been learnt from other individuals; when referring to asocial learning I am including instances of innovation given that definitions of asocial learning and innovation are often hard to tease apart (see Carr, Kendal & Flynn 2016). Moreover, definitions of innovation include that the behaviour is novel in the population (e.g. Reader & Laland 2003), therefore it must have been learned asocially. New behaviours in a population have to be introduced by someone, thus some individuals in a population learn new behaviours through sampling the environment directly, for example via trial and error learning, or modifying an already learnt behaviour. Which individuals tend to perform these novel behaviours has been found to co-vary with factors such as boldness, neophilia and risk-taking (Brosnan & Hopper 2014).

The costs and risks involved in learning asocially or socially will determine when individuals engage in either form of learning. Performing a new behaviour can often be costly or risky to the individual, either because it involves overcoming neophilia to interact with a novel object or because the pay-off is unknown (Brosnan & Hopper 2014). Relying on asocial learning may simply be more costly in terms of time and energy than acquiring the behaviour socially

(Kendal et al. 2005; Arbilly et al. 2011). Social learning can be used to help avoid the associated risks with obtaining information asocially, particularly in changing environments where asocial learning can become unreliable (Boyd & Richardson 1988; Feldman, Aoki & Kumm 1996). In bumblebees, starlings, sticklebacks and frog-eating bats, variable pay-offs in personal information cause an increased reliance on social learning (Rafacz & Templeton 2003; van Bergen, Coolen & Laland 2004; Jones, Ryan, Flores & Page 2013; Smolla, Alem, Chittka & Shultz 2016). Thus, there is within-individual variation in the use of social versus asocial learning; social learning may be employed when the environment is risky and personal information is unreliable in its pay-offs.

Consistent between-individual differences, sometimes termed ‘animal personality’ in the behavioural ecology literature (Eysenck & Eysenck 1985), have also been found to predict innovation rates in a diverse range of non-human animals. For example, differences in traits such as neophilia, exploratory behaviour and boldness predict asocial learning in fish, bird and primate species (e.g. Verbeek et al. 1994; Greenberg 2003; Boogert, Reader & Laland 2006; Tebbich, Stankewitz & Teschke 2011; Guenther et al. 2014). Individual differences in the “boldness shyness continuum” (Wilson et al. 1994), broadly defined as the willingness to take a risk (Smith & Blumstein 2008; Réale et al. 2007) are arguably explained by differences in neophobia, with less neophobic individuals more likely to approach a novel item, and therefore interact with it to solve a problem or retrieve a reward. As well as being more likely to solve the task, bolder individuals are also more likely to solve novel tasks more quickly and efficiently, as was the case in cavies and Darwin’s finches (Guenther et al. 2014; Tebbich, Stankewitz & Teschke 2012). Less neophobic individuals may also be more likely to approach and solve a novel task asocially, as has been shown in starlings, guppies and chimpanzees (Boogert, Reader & Laland 2006; Laland & Reader 1999; Massen et al.

2013). Therefore inter-individual differences in traits such as neophilia and boldness may predict the tendency to use asocial over social information sources.

Differences in traits such as boldness and neophilia can often be understood in terms of risk-taking behaviour (Wolf et al. 2007). Individual differences in risk-taking behaviour can be partly explained through differences in life-history trade-offs, arguably contributing to the presence of sex differences in risk-taking behaviour found in a broad range of species (Wolf et al. 2007; Biro & Stamps 2008; Bengston, Shin & Dornhaus 2016; Orsini & Setlow 2017). For example, males have been found to be bolder, or more aggressive, than females in a range of species such as cichlids, sticklebacks, rats, prairie voles and guppies (for review see Smith & Blumstein 2008). Aggressiveness, or boldness, could evolve in response to competition for mates via sexual selection, and is therefore expected to be found more often in males than females when males are competing for mating opportunities (Trivers 1972). Likewise, for species in which females initiate copulation, higher levels of aggression in females compared to males are found (Cole et al. 1980; Budaev et al. 1999). Sex differences in risk-taking have been linked to decision-making in rats, in which males chose a larger reward associated with a risk of a small shock, instead of a small, safe reward, significantly more often than females did (Orsini et al. 2016). Although risk-taking differences predict differences in asocial learning and decision-making in a range of non-human species, a question remains as to whether sex differences in risk-taking affect men and women's asocial and social learning choices.

Men show higher average risk-taking scores than women on a range of measures (Byrnes, Miller & Schafer 1999; Cross, Copping & Campbell 2011; Charness & Gneezy 2012; Cross, Cyrenne & Brown 2013), although there is large overlap between men and women's scores on many of the risk-taking measures (Nelson 2015) and not all risk-taking measures exhibit

a sex difference in scores (Harris, Jenkings & Glaser 2006). Some researchers have argued that the observed sex differences in risk-taking measures can be explained in terms of evolutionary theory, in that men and women differ in their parental investment with women having more to lose than men in terms of reproductive success by taking physical risks (Daly & Wilson 1988; Campbell 1999). Others highlight the importance of cultural and developmental factors (Nelson 2015). For example, in an experimental lottery and investment task, boys and girls in same-sex education did not show the reported sex difference in risk-taking that co-educated children did (Booth & Nolen 2012). Similarly, men who took part in a same-sex discussion group on risk scenarios gave higher risk-taking scores than controls and their own pre-experiment scores (Ronay & Kim 2006). The authors interpreted this difference in terms of social identity theory (Tajfel 1982); the men were motivated to conform to the cultural norms and expectations of their gender. Finally, sex differences in risk-taking measures are not consistent cross-culturally; participants including an African Masaai group and South Asian Khaai group did not show a sex difference in risk-taking measures (Gneezy et al. 2009). Some research suggests that the largest difference in risk-taking measures is not between men and women, but white males and everyone else (Kahan et al. 2007). It is argued that white males perceive less risk on average compared to non-white males and females due to their position at the top of a social hierarchy (Kahan et al. 2007; Flynn et al. 1994). Thus, there are many potential factors involved that contribute to apparent sex differences in risk-taking measures.

Regardless of the evolutionary, developmental or cultural factors that may contribute to the presence of sex differences in some risk-taking measures, the impact of these sex differences on the tendency to use social or asocial information has not yet been explicitly addressed. This is the basis of the experimental work in **Chapter 2,** in which I investigate

whether varying the riskiness of social and asocial sources of information leads men and women to choose different information sources on a novel computer-based task.

1.3.4 Sex Differences in Asocial Learning

Sex differences in rates of innovation have been found in a range of non-human animal species such as meerkats, lemurs, guppies and primates (Thornton & Samson 2012; Dean, Hoppitt, Laland & Kendal 2011; Laland & Reader 1999, Reader & Laland 2001). Nevertheless, the degree and direction of the sex difference, as well as the postulated explanations, are thus far variable and mixed. For example, Thornton and Samson (2012) gave extractive foraging tasks to a group of wild meerkats and found that the majority of the innovators were males. These authors suggested that males, which are the dispersing sex, might be more innovative than females due to their need to explore and learn in novel environments. Likewise, an experiment with seven groups of ruffed lemurs of varying sex ratios, found that individuals of the less represented sex showed higher rates of innovation than individuals of the other sex when given a novel problem to solve (Dean et al. 2011). Here, the authors proposed that the sex difference could be explained by female dominance found in this species, with females having greater access than males to novel food sources. An experimental study found that foraging innovation was positively correlated with competitive ability in male, but not female, guppies (Laland & Reader 1999). The proposed explanation involved differing levels of parental investment between the sexes, and the authors predicted greater levels of foraging innovation in females when maternal investment exceeds paternal investment. In contrast to these predictions, a large meta-analysis looking at 22 species of wild and captive non-human primates found that males innovated more than females (Reader & Laland 2001). This pattern persisted when controlling for research effort and when the chimpanzee data were analysed separately, as most innovation observations

were of chimpanzees (Reader & Laland 2001). Thus, there appear to be sex differences in innovation rates in a range of species, although the underlying mechanisms and factors that contribute to these differences are poorly understood.

The extent to which sex differences are found in the tendency to solve a novel problem individually, or to innovate, is so far unclear and is likely to be intertwined with the specific life-history traits, ecology, social structure, and developmental environment of the species and individuals (Smith & Blumstein 2008). Indeed, it has been argued that sex should be viewed as a biological factor to act as a proxy for other, unknown factors that may explain variation in behaviour better than sex (Maney 2016). It may be the case that sex differences actually reflect individual differences in traits such as boldness and aggression, as mentioned above. Similarly, sex differences may act as a proxy for hormonal differences which explain variation in behaviour more accurately then just the sex of the individual. A more informative and accurate approach to apparent sex differences in behaviour is to ask how much variation is accounted for by sex, and to what extent the sexes differ, rather than whether or not sex "explains" the difference (Maney 2016). Accordingly, I am interested in looking for alternative factors that help to explain apparent sex differences in behaviour, such as individual differences in risk-taking tendencies.

1.3.5 Sex Differences in Social Learning

Few animal studies have examined sex differences in social learning explicitly. One such study comes from observations of wild chimpanzees using tools to fish for termites (Lonsdorf et al. 2004). Based on a four-year longitudinal study of 14 chimpanzees under 11 years old, the researchers found that female chimpanzees fished more often and from an earlier age than males. The study found that the females' technique (i.e. length of stick and number of dips) matched the mothers' more closely than the males' technique did, and that the females

were more proficient at fishing for termites than the males were (i.e. the number of termites gathered per dip). Despite the fact that there was no difference found in the time the mother spent with their infants, or the tolerance for their infants between the sexes, females watched their mothers more, while males played at the mound more than individuals of the other sex. However, this study was based on a small number of chimpanzees and has yet to be replicated directly. Furthermore, no sex difference was found in equivalent measures and observations of ant-dipping behaviour (Humle, Snowdon, Matsuzawa 2009).

Zefferman (2016) has argued that sex-biased transmission of tool use seen in dolphins (Mann & Sargeant 2003) and chimpanzees (Lonsdorf et al. 2004) could theoretically be explained via uniparental teaching. Using a mathematical model, Zefferman argued that mothers are predicted to invest more in teaching daughters, who then go on to teach granddaughters, than sons, as it leads to larger inclusive fitness benefits. However, the sex biased transmission of tool use that he refers to has only been reported in dolphins and chimpanzees (Mann & Sargeant 2003; Lonsdorf et al. 2004). Although these two isolated examples of sex-biased tool use transmission are interesting and worth understanding, it is unlikely that they represent a general adaptation to uniparental teaching that would be generalisable to many other species as stated. A recent study of captive bonobos found that females were quicker to attempt and to solve a termite fishing problem (Boose, White & Meinelt 2013), however there is no way of knowing from the task whether the bonobos acquired the tool use socially or asocially. Furthermore, any sex differences in social learning of tool use in chimpanzees and bonobos likely reflects the sex difference in tool use generally in these species (McGrew 1975; Gruber, Clay & Zuberbuehler 2010).

One study looking at the social transmission of tool use in children reported a sex difference in which more girls than boys “dropped out” of the task due to a failure to be an effective

model for the next participant, and boys gained higher achievement scores than girls in their first demonstration sessions (Flynn & Whiten 2008). However, the higher drop-out rate in girls, as well as the higher achievement score in boys, only occurred on the first demonstration session, and not on either the first two solo attempts or the second demonstration. Furthermore, the higher achievement score in boys was only present for three year olds and not five year olds. Therefore, the reasons behind this sex difference remain unclear and may be linked to motivation or confidence, as it is only apparent when the participant is first watched by the next participant. In addition, the task involves retrieving a reward from a restricted space inside a puzzle box using a tool, skills that rely on spatial and motor abilities, in which males have been found to generally outperform females (Voyer et al. 1995; Vederhus & Krekling 1996; Davies & Rose 2000; Dorfberger et al. 2009). Indeed many social learning tasks require the learning of new motor skills, making it difficult to tease apart a male advantage in social learning generally, or motor skills specifically. For example, in another puzzle-box based task, Rawlings and colleagues (In Prep) found that boys aged 7-11 years old were more likely to choose to solve a puzzle box asocially than girls were, and girls were more likely to choose to learn socially than boys were.

A task with adult humans by Mesoudi and colleagues (2015) found that females used social learning more than males did in a task involving the virtual construction of arrowheads, although the effect was much smaller than the effect of culture (i.e. participants of East Asian cultures used social learning more than those of Western cultures), which was the focus of the study. Moreover, this finding was not found in other experiments using the same task (Mesoudi & O'Brien 2008; Mesoudi 2011; Atkisson, O'Brien & Mesoudi 2012). Studies explicitly examining social learning in children and adult humans often don't report or analyse sex differences, either because the researchers don't have an appropriate sample of participants, or because it wasn't a variable of interest (e.g. Flynn, Turner & Giraldeau

2016). In some cases, studies exploring social and asocial learning mechanisms have only tested participants of one sex (e.g. Derex et al. 2015). Therefore most claims of sex differences in social learning come from the social psychology literature, which reports that women conform more than men (e.g. Cooper 1979; Eagly & Carli 1981; Bond & Smith 1996). However, the use of the term “conformity” in social psychology research differs to that used in the social learning literature in cultural evolution research, as discussed in detail in the following section.

1.4 Conformity and Confidence

1.4.1 Defining Conformity

Conformity has been defined by cultural evolutionists as a form of frequency dependent social learning bias in which individuals are disproportionately likely to copy the most common behaviour in their group (Boyd & Richerson 1985, p.206). Theoretical models have shown conformity to be more adaptive than copying at random, as it allows individuals to learn the adaptive behaviour of the local group when moving to new environments (Boyd & Richerson 1985; Henrich & Boyd 1998). Most experimental evidence for conformity in humans, however, comes from the social psychology literature, where participants have been shown to override their own personal perceptions to go with a conflicting, and wrong, majority view provided by confederates of the experiment (e.g. Sherif 1936; Asch 1951). This particular conformity effect has been replicated many times since (e.g. Deutsch & Gerard 1955; Latane & Wolf 1981; Baron et al 1996; Bond & Smith 1996; Bond 2005). Social psychologists have typically studied when individuals dismiss their own judgement in the face of ambiguous, or even conspicuously incorrect, information. This is in contrast to how cultural evolutionists have characterised conformity, perhaps because the fields are asking quite different questions. Cultural evolutionists tend to be interested in how behaviours such as conformity evolve in the first place, and what the population-level consequences of conformity are. Social psychologists are typically interested in individual-level conformity, i.e. in what contexts is an individual's behaviour altered by that of other group members, especially when the group's behaviour is in direct contrast to their own perception or knowledge.

One study by Jacobs and Campbell (1961) appeared to bridge the gap between social psychology and cultural evolution experiments by running a transmission chain experiment

with Sherif's (1936) autokinetic conformity task. This task involves judging the distance that a speck of light has moved, and includes confederates exaggerating the distance, similarly to the Asch test in which the judgement of a line length is exaggerated. The researchers wanted to know whether the conformity effect would persist over many generations of individuals when each individual is replaced with a new, naive participant. This transmission chain design is now commonly used by cultural evolutionists to look at how between-group variation and within-group similarities emerge experimentally (Mesoudi & Whiten 2008). Jacobs and Campbell (1961) found that once all original individuals had been replaced with naive individuals, the judgements reverted to those of individual controls. This result contradicts findings of similar studies by cultural evolutionists, but a possible explanation could be that in this case, the behaviour of the rest of the group was in direct conflict with the perceptions of the individual. Indeed, social learning and conformity is adaptive when individual or asocial learning is unreliable (Kendal et al. 2005).

Recent debates around the current evidence for conformist behaviour in both animal populations and human experiments have highlighted the need to distinguish between population-level conformity and individual-level conformity biases (Acerbi et al. 2016). A particular point of contention is that the sigmoidal (or "s-shaped") curve is often used to support claims of both conformist-biased transmission in a population and a conformity-bias in individuals (e.g. Efferson et al. 2008; Mesoudi & Lycett 2009; Morgan et al. 2012). However, as shown through various simulations, the sigmoidal curve can be achieved through other forms of learning bias and does not necessarily point to the presence of a conformity bias (Acerbi et al 2016). An alternative way that the sigmoidal curve can be generated is if individuals are copying a subset of the population, such as dominant, popular, or more conspicuous individuals. Another way that the sigmoid curve would be present is if individuals had an underlying bias for one of the two alternative behaviours (Acerbi et al

2016). Thus, the presence of a sigmoidal pattern in the adoption of majority behaviours does not always bolster a case for conformity.

Another distinction that needs to be clearer is between when an individual has observed a majority of individuals display a particular behaviour and when an individual has observed a majority of behaviours being displayed by a smaller number of individuals (van Leeuwen et al. 2016). For example, a chimpanzee may view three conspecifics each press a particular button once, whereas another chimpanzee may see just one conspecific press that button three times. It has been argued that some claims of conformity, particularly in animal observations, could be explained by animals simply observing a behaviour more often than other behaviours, rather than an underlying bias for copying a majority of individuals (van Leeuwen et al. 2016). Many of the discussed criticisms for claims of conformity are particularly relevant to animal studies and indeed have been aimed at two studies in particular with vervet monkeys and great tits (van de Waal, Borgeaud & Whiten 2013; Aplin et al. 2015). Deciphering which behavioural strategies animals are using often relies upon making inferences based on their behaviour in naturalistic experimental settings. However, with human studies, experiments can be more precise and controlled due to the use of computer-based tasks, and due to the benefit of being able to use language to communicate with the participants. Thus, some of the discussed criticisms that are directed at animal studies of conformity are less relevant for human experimental work on conformity.

Convincing evidence for individual-level conformity biases in humans has been put forward by Morgan and colleagues (2012) who ran several computer-based experiments exploring the contexts in which people use social learning. One of these experiments involved participants completing 25 trials of two-alternative forced-choice mental rotation tasks. After each trial, participants recorded their confidence in their answer, and were then shown what

a group of other participants, or “demonstrators”, thought the correct answer was. The number of demonstrators varied from 5 to 10, and the proportion of demonstrators agreeing and/or disagreeing with the participant varied. Morgan and colleagues (2012) found that participants did indeed display individual-level conformity biases, by switching their answer to match the majority’s answer disproportionately more often than chance, given the proportion of demonstrators disagreeing with them. This experimental design avoided some of the criticisms of other conformity studies. Firstly, the participants gave their own answer first, without seeing input from anyone else, so that the use of social information is in direct comparison to what participants did individually. Secondly, participants see a varied number of demonstrators, with varying proportions that disagree with them, rather than just a unanimous opposition. Finally, as participants were only exposed to flashing tiles that represented demonstrators’ answers, and not the demonstrators themselves, it is unlikely that participants could have been copying a subset of the demonstrators. Therefore, it seems reasonable to conclude that participants were using a copy-the-majority social learning strategy in this task.

1.4.2 Sex Differences in Conformity

In classic conformity experiments in the field of social psychology, such as the Asch line test mentioned above, women have been found to conform more than men (e.g. Cooper 1979; Eagly & Carli 1981; Bond & Smith 1996). For example a meta-analysis by Bond and Smith (1996) looked at 133 studies of the Asch experiment from 17 countries and found that the percentage of female participants had the largest impact on the conformity effect of all the moderator variables. However, information on the sex of participants wasn’t present for one sixth of studies, and for those in which it was, more than half of them used all men, and a quarter used all women. Nevertheless, sex was still found to have a significant moderating

effect when comparing male-only studies to female-only studies, as well as looking within studies that contained both men and women. A recent replication of the Asch task used a visual illusion to display the stimuli, so that participants unknowingly saw different line lengths from each other and no confederates were needed (Mori & Arai 2010). In this version of the task, only the females conformed to the opposing majority, whilst males did not. However, participants were tested in same-sex groups of 4 at a time, and only 26 out of the 164 participants saw a majority disagree with them. Therefore, the number of participants who could have conformed to a majority in this study was small, and the reasons for the apparent sex difference in conformity are speculative and unclear.

One factor that appears to moderate the sex difference in conformity is anonymity; sex differences in conformity are reportedly smaller when the participants in the experiment can conform anonymously (Eagly & Carli 1981). One study found that when responses were overseen, men behaved in a non-conformist way whereas women's conformity levels stayed the same as in the anonymous condition (Eagly, Wood & Fishbaugh 1981). The authors suggest that rather than women behaving in a conformist way, men actually behave in a non-conformist way when responses are visible to others. Recent research in computer mediated communication, such as through emails or online forums, have had mixed results in terms of sex differences in conformity and how it is affected by anonymity. For example, Guadagno and Cialdini (2002) looked at dyadic online interactions and found that prior interaction was related to higher conformity in women but not in men. In contrast, Lee (2006) found that women were more likely to conform to a group decision, but only when in a depersonalised context, i.e. lacking any individuating or personal information about each other. A follow-up study by Guadagno and Cialdini (2007) found that men were more easily persuaded than women, but only when shared identity with the persuader was low. Finally, a study exploring group identity used participants from various online communities and forums

and gave them a range of questions on logic, attitudes and general knowledge (Rosander & Eriksson 2012). Some participants were shown answers from members of their online community in which the majority gave the wrong answer. In contrast to previous findings, men conformed more than women in both the logic type of questions and in the more difficult questions (Rosander & Eriksson 2012). Thus, there are many potential variables that influence sex differences in conformity and the effects of being observed by others in the group.

One relevant factor that could explain some of the sex differences in conformity is differences in confidence. Morgan and colleagues (2012) ran a variety of experiments looking at when participants utilised social information to solve tasks as mentioned above. After seeing the demonstrators' answers, participants were then given the opportunity to change their mind or stick with their initial answer. Participants were more likely to switch to a majority's answer when their confidence was low. Similarly, Cross and colleagues (2016) found that women's confidence on a mental rotation task was lower than that of men's; lower confidence predicted higher conformity, and sex had an indirect effect on conformity mediated by confidence. In a comparable letter-transformation task, in which men and women's confidence levels did not differ, there was no indirect effect of sex on conformity (Cross et al. 2016). One explanation for the confidence differences on the different tasks could be due to stereotypes about the performance of men and women on mental rotation tasks. Indeed, early studies comparing gender-stereotyped questionnaire items found that women were more likely to conform to a majority's answer when the content was stereotypically masculine, such as fishing, whereas men were more likely to conform to a stereotypically feminine item, such as knitting (Goldberg 1975). Further evidence of sex differences in confidence are reviewed in the following section.

1.4.3 Sex Differences in Confidence

Women are reported to have lower self-confidence than men in a broad range of areas, both in cognitive tasks, within the work-place and throughout education (e.g. Syzmanowicz & Furnham 2011; Blanch, Hall, Roter & Frankel 2008). A study involving meta-analyses of gender differences in confidence in mathematical, spatial, verbal and general intelligence tasks found that women were less confident than men in all except verbal tasks (Syzmanowicz & Furnham 2011). Males rate their intelligence higher than females do (Furnham 2001), and this has been found to be the case across 12 different nations (Von Stumm, Chamorro-Premuzic & Furnham 2009). Both boys and men are more self-confident than girls and women in their writing ability (Pajares, Miller & Johnson 1999; Pajares & Valiante 2001). A study of Swedish school children found boys were overconfident in their maths ability, while girls were underconfident (Dahlbom, Jakobsson, Jakobsson & Kotsadam 2011), a finding that has been replicated in El Salvadorian children also (Jakobsson, Levin & Kotsadam 2013). Within higher education, women underrate their performance in macroeconomics (Jakobsson 2012) and biology (Langan et al. 2008). The same pattern is found in medicine, despite empirical evidence showing their performance is equivalent, if not superior to, their male peers (Blanch, Hall, Roter & Frankel 2008).

Sex differences in confidence have been found in mental rotation tasks (Cooke-Simpson & Voyer 2007), in which sex differences in general have been studied extensively (e.g. Reilly & Neumann 2013). Mental rotation tasks typically display four arrangements of cubes in various configurations, and participants have to pick the two that match an example shape. Two out of the four shapes have been rotated, the other two are different shapes entirely (Vandenburg & Kuse 1978). Another version typically used is that of Shepard & Metzler (1971) version in which only two shapes are displayed and participants have to state

whether they match or not. These tasks seem to produce a reliable sex difference in that men outperform women on average, with a meta-analysis finding an overall effect size of *d = 0.67* (Voyer et al. 1995). Many potential explanations for this sex difference have been put forward, from evolutionary psychology perspectives (e.g. Eals & Silverman 1994; Silverman, Choi & Peters 2007), developmental perspectives (e.g. Feng, Spence & Pratt 2007; Quaiser-Pohl, Geiser & Lehmann 2006), hormonal differences (Clint et al. 2012; Aleman et al. 2004) and gender roles in society (Reilly & Neumann 2013). However, the magnitude of the sex difference, and indeed its presence at all, varies depending on the type of stimuli used in the mental rotation (Voyer et al. 1995; Grimshaw et al. 1995; Voyer & Hou 2006; Alexander & Evardone 2008). There are potentially many factors that affect men and women's performance on mental rotation tasks; one of these factors could be underlying differences in confidence, as discussed below.

Cooke-Simpson & Voyer (2007) found that men were significantly more confident in their mental rotation responses than women were. Likewise, Cross and colleagues (2016) found that women's confidence ratings were lower than expected for their performance in a mental rotation task, but not in a letter-transformation task. Estes & Felker (2012) explicitly addressed the role of confidence in affecting mental rotation performance. Confidence predicted performance both between and within sex, and manipulating participants' confidence by giving them false feedback on a previous line-judgement task significantly affected performance. In conditions in which participants weren't allowed to omit responses, no sex difference in performance was found (Estes & Felker 2012). The authors argued that because participants had to answer every trial in these conditions, evaluating confidence on each trial was ineffective and thus irrelevant to the task. By making confidence irrelevant to the task in this way, sex differences in performance on the task were eliminated. However, other studies of the mental rotation task in which participants have to provide an answer on

every trial, as well as their confidence, found sex differences in performance on the task (Morgan et al. 2012; Cross et al. 2016). Therefore, the impact of making every trial compulsory on participants' confidence and performance is unclear. Nevertheless, it appears that confidence may play an important role in sex differences on mental rotation tasks.

1.4.4 Confidence and Stereotype Threat

Sex differences in confidence may be partly due to the phenomenon of stereotype threat. Stereotype threat "refers to the conscious or unconscious belief that one belongs to a group stereotypically known for specific performance deficits... these beliefs lead to suboptimal performance on tasks specific to the stereotype," (Stoet & Geary 2012). The stereotype threat effect has frequently been explored in the context of mental rotation tasks due to the assumption that awareness of a sex difference in this task might contribute to the sex difference in performance (Moé & Pazzaglia 2006). By applying stereotype threat, or conversely stereotype lift (in which attempts at improving the affected group's performance are made), the sex difference has been increased or reduced respectively. For example, Sharps, Price and Williams (1994) found that when the spatial elements of the task were emphasised in the instructions, boys outperformed girls as expected, but when the spatial element of the task was de-emphasised, boys did not outperform girls. Another study primed female and male participants with descriptions of a "typical day in the life" of a female or male character before taking the test (Ortner & Sieverding 2008). The descriptions contained gender stereotyped words to activate either a female or male stereotype. Both sexes performed worse in the female stereotype condition, and sex differences in performance disappeared when women were primed as male. Similarly, students at a private college were primed to either think about their status as a student at a selective college or as a student of their particular sex (McGlone & Aronson 2006). Females performed worse on a mental

rotation task when primed to think about their sex compared to when primed to think about their student status, and the reverse was true for male students.

The mechanism underlying the stereotype threat phenomenon has yet to be deciphered, with some suggesting that stereotype threat causes a decrease in performance by decreasing motivation (Steele & Aronson 1995; Walton & Cohen 2003), whilst others suggest that it affects performance by increasing cognitive load (Croizet et al 2004; Schneeberger & Williams 2003). Clearly confidence is important, and no one has yet explicitly tested whether stereotype threat affects confidence. Previous research shows that confidence affects mental rotation performance (Estes & Felker 2012), and that stereotype threat affects mental rotation performance (e.g. McGlone & Aronson 2006; Moé & Pazzaglia 2006; Hausmann, Schoofs, Rosenthal & Jordan 2009; Heil, Jansen, Quaiser-Pohl & Neuburger 2012), but whether confidence is the underlying mechanism by which these effects occur remains to be determined. This research question is addressed by experimental work in **Chapter 3** in which I measure men and women's confidence in a mental rotation task under different conditions involving stereotype threat primes. I also wanted to know whether manipulating men and women's confidence using stereotype threat would have an effect on their likelihood to conform in the task. Thus I not only measure participants' confidence levels in response to stereotype threat conditions, but also measure the impact that participants' confidence has on the likelihood to adopt a majority decision in the task. However, the reliability of the stereotype threat effect itself has come under question as a number of meta-analyses have suggested that the literature may be suffering from publication bias. This uncertainty surrounding stereotype threat as well as other forms of priming and the replicability of psychology studies in general will be discussed further in the next section.

1.5 Uncertainty in the Stereotype Threat Effect

1.5.1 The Stereotype Threat Literature

There is a large volume of literature on stereotype threat, with over one hundred papers examining the effect of stereotype threat on women's maths, science and spatial skills alone (Flore & Wicherts 2015), and seven independent meta-analyses on various aspects of those reported effects (Walton & Cohen 2003; Nguyen & Ryan 2008; Walton & Spencer 2009; Stoet & Geary 2012; Picho, Rodriguez & Finnie 2013; Flore & Wicherts 2014; Doyle & Voyer 2016). Although six of these meta-analyses reported significant effects of stereotype threat on either maths or spatial skills, the effect sizes varied from small to medium, and four of the meta-analyses explicitly stated that publication bias is likely to be distorting the literature considerably. Flore and Wicherts (2015) argued that the stereotype threat literature is particularly vulnerable to publication bias as it meets criteria laid out in Ioannidis' (2005) influential paper detailing how most published scientific findings are false. Ioannidis argued that a field will be particularly susceptible to publication bias if it involves small sample sizes, small effect sizes and if it allows flexibility in definition, study design and outcome measures. Stereotype threat has, like many scientific phenomena, many definitions. Moreover, stereotype threat is particularly difficult to conceptualise, as it relies on a pervasive belief within a culture or society, yet is also "activated" within experimental contexts. It appears that some studies attempt to activate a stereotype threat that they assume to be already present in participants, whereas others try to explicitly create a stereotype within an experimental context. The concept is confused further by the fact that some studies, either intentionally or incidentally, cause a stereotype "lift" or "boost" effect in which the non-negatively stereotyped group experiences a relative increase in performance (Walton & Cohen 2003; Hirnstein et al. 2014). Not having a clear or precise description of a) what the phenomenon consists of, b) in

what contexts it is expected to be found, and c) predictions of how it could be experimentally induced, manipulated and measured, means that the stereotype threat literature is inconsistent, and most likely subject to publication bias.

1.5.2 The Replication Crisis in Psychology

The stereotype threat literature can be understood as part of a wider priming literature in psychology in which participants are given information prior to, or during an experiment, and this information has a detectable effect on their behaviour. However, whilst priming has been reported to affect a wide range of behaviours such as performance on intelligence tests, decision making, individualism and collectivism and walking speed (Bower 2012; Doyen et al. 2012), the field has achieved little consensus with regards to the genuine existence of such effects. Concern over the lack of robustness in psychological findings, in general, led to the Reproducibility Project in which a global collaboration of scientists attempted to replicate 100 findings from three top psychology journals published in 2008 (Open Science Collaboration 2015). The project only managed to successfully replicate the results of 30 of the original findings (but see critique by Gilbert et al. 2016). Since then, many more failed replication attempts have come to light and gained wide-spread attention not only from the academic community but also in the media (e.g. Ioannidis 2015; Hagger et al. 2016; Ranehill et al. 2015; Smith 2016).

Although the field of psychology has received particular attention, evidence of replication failures and publication bias are present in many, possibly all, areas of science (e.g. Horton 2015; Ioannidis 2005; Enserink 2012; Peterson 2016; Vul, Harris, Winkielman & Pashler 2009) and psychology has in fact been heralded as setting a precedent for explicitly addressing the problem in the field (Gelman 2016). The replication crisis in psychology is

part of a larger debate surrounding the current process of scientific publication in general, including problems such as publication bias, p-hacking, the misuse of statistical techniques and the “file-drawer” effect, problems that have existed in science for many years with mixed opinions and widespread debate (e.g. Bissell 2013; Bohannon 2014; Kahneman 2014; Schnall 2014; Fischer 2015; Pulverer 2015). Recent proposals such as guidelines against the mis-use of p-values (Wasserstein & Lazar 2016), banning the p-value (Trafimow & Marks 2015), pre-registering methods (Chambers et al. 2014; van’t Veer & Giner-Sorolla 2016), and the Open Science movement (e.g. Kidwell et al. 2016) are propagating discussion and endorsement of substantial changes to scientific publishing and research methods.

Recently, researchers have modelled the scientific process itself in order to establish which changes are likely to have the most positive outcomes (McElreath & Smaldino 2015). In this way, hypothetical scenarios, such as researchers only publishing novel positive findings, can be explored and meaningful implications from the model can be discussed (McElreath & Smaldino 2015; Rzhetsky et al. 2015; Smaldino & McElreath 2016). For example, McElreath and Smaldino (2015) created a mathematical model of scientific discovery to simultaneously examine the effects of hypothesis formation, replication, publication bias and variation in research quality. They found that replication studies help to gradually decipher true hypotheses from false, but are just as vulnerable to problems such as false positive rates. Additionally, the communication of negative replications is more beneficial to the scientific process than the communication of novel negative findings. Complementing this study, Smaldino and McElreath (2016) created an evolutionary model of science representing labs competing for publications and consequently prestige and jobs. Members of labs inherit their methods culturally, thus more successful labs’ methods ‘survive’. The model demonstrated that, due to incentives for high numbers of publications, poorer methods that produced false positives persisted and proliferated. Although replication slowed this process down, it could

not halt it altogether; hence the authors concluded that institutional changes of incentives are required to improve the quality of scientific research.

1.5.3 Frequentist versus Bayesian Approaches to Statistical Analysis

There are many possible improvements that could be made to the process of scientific publishing; one area of improvement that is often focussed on is the appropriate use of statistical analyses. Indeed, discussions around the recent increase in statistical errors being found and publicly exposed from peer-reviewed psychology papers reached a climax with an exchange between psychologist Susan Fiske and statistician Andrew Gelman (Gelman 2016; Fiske 2016). Spurred on by this debate, many psychologists are endorsing the exposure of statistical mistakes as a means of improving statistical practice in the field. Notwithstanding erroneous statistical practices, the use of null-hypothesis significance testing (NHST) in general has come under criticism and is often contrasted with Bayesian approaches as an alternative (e.g. Kruschke 2011; van de Schoot 2012; Lee & Wagenmakers 2005). Many advocates of Bayesian methods of data analysis have previously tried to introduce the approach to psychologists as an alternative to classical frequentist methods (e.g. Rouder et al. 2012, van de Schoot et al. 2014; Yuan & Mackinnon 2009), but without widespread success or response from the field.

It is argued that Bayesian approaches have many advantages over frequentist methods, such as the flexibility to create highly complex models tailored specifically to your dataset, better handling of non-normal parameters, better handling of small sample sizes, a more intuitive interpretation of results than significance testing, the ability to accumulate and update knowledge based on previous research, and a more explicit inclusion of researcher

uncertainty (van de Schoot et al. 2014; Bolker et al. 2009). Bayesian inference can also solve the multiple comparisons problem that occurs when conducting multiple statistical tests, so common in psychology research (Gelman & Tuerlinckx 2000). Gelman and Tuerlinckx (2000) demonstrate that Bayesian inference is more conservative than frequentist inference in that it is consistent with previous inferences unless the data forces a change (Gelman 2010). Therefore, if Bayesian methods are more conservative, they may be of use for helping to minimise false positive findings.

Use of 'priors', at the most obvious level, can be used to restrict the search space to theoretically plausible values and prevent exploration of space that is logically impossible. To give a very crude example, if the variable of interest is body temperature, a prior can be set for the reasonable values of body temperature (e.g. between 36 and 38 degrees centigrade, with room for error), so that values outside of these ranges are not explored unnecessarily. Additionally, "weakly regularizing" priors can be used to help prevent a model overfitting to the data (McElreath 2016). Ideally, priors are tuned to avoid both overfitting and underfitting; the model can learn about the patterns in the data, but doesn't place so much weight on the data that it is no longer generalisable to other samples. Thus, employing appropriate priors can help researchers to make the most of their data in terms of uncovering patterns, but also making meaningful and realistic out-of-sample predictions.

Of particular interest is the potential use of "posterior passing" where results from one experiment are used to inform the analysis of data collected in another (Beppu & Griffiths 2009). Proponents of posterior passing argue that it could have strong implications for how scientific fields develop certainty about supposed effects. The posterior passing approach was used by Beppu and Griffiths (2009) and led to successively better inferences over time in a lab experiment and a theoretical model of simulated participants. These researchers

concluded that information about the certainty and plausibility of information, as well as the observations themselves, are needed for knowledge and inference to cumulatively improve over time. This same philosophy can be applied when thinking of science as a whole, especially in the context of the replication crisis. To avoid each experimental dataset carrying too much weight, priors could be used to incorporate what is already known, or what the data are expected to look like, based on previous studies. As long as priors allow room for error, the model will be predominantly directed by the data, but not to the extent that each dataset is treated as a completely isolated case outside of the context of the rest of the relevant data in the literature.

Thus far, psychologists employing Bayesian methods of analysis have largely used minimally informative priors. Indeed Bayesian methods have often been criticised and misunderstood as allowing any results to be found by using priors to alter the results of an analysis. However, these criticisms are unfounded and stem from a misunderstanding of Bayesian methods and the use of priors. Nevertheless, the use of informative priors to analyse experimental datasets in psychology has not yet been explored. Bayesian methods have only recently started to be adopted by researchers in areas such as psychology, evolutionary biology and neuroscience, and so the implications of implementing priors, other than minimally informative ones, are not currently understood. Therefore, in **Chapter 4,** I present a model which compares frequentist and Bayesian methods of analysis over many datasets simulated to represent the stereotype threat literature. The Bayesian methods include a form of “posterior passing” in which the estimates from a previous analysis are used as priors for the next analysis. This use of priors was implemented to explore whether it would increase certainty about the presence or absence of an effect over many replicated studies, such as is the case in stereotype threat literature.

Thesis Outline

In this thesis I address some of the questions raised by the literature discussed above. Specifically, in **Chapter 2,** I address the question of whether altering the riskiness of social and asocial information differentially affects men and women's tendency to use these information sources. Specifically, I wanted to know whether men would choose to learn asocially more than women when asocial learning is made to be riskier than social learning. I present a novel computer-based experiment that involves constructing virtual spaceships. Participants were given the choice between using social or asocial sources of information, and I manipulated how risky these sources of information appeared to the participants. In **Chapter 3,** I provide more evidence of a relationship between confidence and conformity in an experimental setting using a novel version of a mental rotation task in which 3D images of human figures were used rather than blocks of cubes. I also tried to experimentally manipulate confidence in this task using the stereotype threat paradigm. In **Chapter 4**, I simulated 60 datasets to represent multiple typical stereotype threat priming studies, and analysed these datasets using five different methods of analysis, i.e. commonly used frequentist methods, and more novel forms of Bayesian analysis that employ informative priors based on parameter estimates of previous analyses. I predicted that using informative priors would lead to more certainty about the presence or absence of an effect over many datasets. In **Chapter 5,** I discuss the implications of my findings within the context of cultural evolution research in general, as well as the implications for improving research methods and statistical analysis in psychology.

Chapter 2:

Risk-taking and Social Learning

2.1 Introduction

This study aims to address whether altering the riskiness of social and asocial information will differentially affect men and women's tendency to use these information sources. Individuals use social and asocial learning to differing extents depending on the trade-offs between the reliability and cost of information (Kendal et al. 2005). Individual differences in risk-taking predict differential use of social and asocial learning in a range of non-human species (e.g. Brosnan & Hopper 2014). In humans, men give higher average scores on risk-taking measures than women (e.g. Cross, Copping & Campbell 2011). Thus, I hypothesised that when asocial learning is riskier than social learning, women will rely more heavily on social learning than men will, due to their lower levels of risk-taking compared to men. To test this hypothesis I created a novel computer-based task in which participants had to construct a virtual spaceship then choose between using social or asocial sources of information to complete the task. The spaceship task was chosen to be novel and opaque to the participants, so that they were unlikely to have any previous knowledge relevant for completing the task. The task was designed in a similar vein to other computer-based tasks used to study social learning strategies such as the arrowheads task (Mesoudi & O'Brien 2008), the fishing-net task (Derex, Godelle & Raymond 2012), and the "Creature League" (Wisdom & Goldstone 2010) in that participants had to make decisions on multiple elements to achieve an overall score. At the end of the game, participants were also given a risky impulsivity personality questionnaire (Campbell & Muncer 2009). This was to check if the sample of participants displayed the predicted sex difference in risky impulsivity, in that men are found to have higher risky-impulsivity scores than women on average.

Participants were incentivised with a monetary bonus to try and achieve the top score for their spaceship. I manipulated the riskiness of the social and asocial information in each condition; in Condition 1 ("Social Risky"), choosing to use social information was risky but

choosing to use asocial information was comparatively safe. In Condition 2 (“Asocial Risky”), choosing to use asocial information was risky but choosing to use social information was comparatively safe. In the Control condition both options were safe. Social information consisted of viewing other participants’ spaceships and being able to copy some items from their spaceship. Asocial information consisted of viewing new spaceship items not previously seen by the participants. When a source of information was risky, participants were told that if they chose it then their score could greatly increase or decrease. When a source of information was safe, participants were told that if they chose it then their score was guaranteed to remain the same, or increase slightly. The control condition was implemented to detect whether participants had a preference for social or asocial learning when both options were safe, and to compare this behaviour to participants in the other two conditions.

The aim of this study was to examine whether altering the potential payoffs of using asocial and social sources of information would differentially influence the probability that men and women used these sources. Here I am defining riskiness in terms of variation in expected payoffs (Wang 2002). I predicted that, when one of these sources of information appeared to be risky, women would be more likely than men to use the alternative source of information. In the control condition, no sex difference in the use of asocial and social information was predicted.

Although not the primary focus of the study, I wanted to ensure that participants were motivated to achieve the top score in the study, and were thus behaving in a way consistent with the “Bounded Risk Distribution” model in which individuals are expected to maximize their probability of reaching a certain goal while minimizing their probability of falling below a certain threshold (Wang 2002). In this instance, participants are trying to achieve the goal of a top score, and theoretically want to minimize their chance of falling below this threshold in

order to achieve a bonus payment. Thus I predicted that when participants were the lowest rank, they would be more likely to choose the risky option.

2.2 Methods

2.2.1 Participants

88 participants (50 females and 38 males) were recruited through the University of St Andrews online recruitment system (SONA). All participants were over 16 years of age, with the majority falling into the 18-25 age range (91%). Participants read an information page and gave their full consent before taking part in the experiment. Participants were given the opportunity to ask the experimenter questions before and after the experiment, and the experimental protocol was approved by the University of St Andrews ethics committee. All participants were reimbursed £3 for their time and were given the opportunity of winning an additional £2 during the experiment. Participants were randomly assigned to three conditions (Control = 13m, 15f; Asocial Risky = 10m, 18f; Social Risky = 15m, 17f) during the experiment as described below and were tested in groups of 4 - 9 people at a time to give the impression of competing for the top score within a group. At the start of the task, participants provided their age, gender, education level and country of origin.

2.2.2 Procedure

The experiment was conducted via an interactive computer game, programmed by Ken Munro and me in the School of Psychology using JavaScript, jQuery, HTML and CSS. The design, appearance and logistics of the game were devised by me. The aim of the game was to construct a virtual spaceship with the highest score out of all players. Players were told within the task advertisement, and verbally at the start of the task, that if they scored the top score they would receive a bonus payment. Players had an initial attempt at building a ship on their own to familiarise themselves with the task, and were given a score and rank

(either 1st, 3rd or 5th) for their first attempt (Phase I). The participants were unaware that their rank and score was randomly assigned, but the task instructions implied that it reflected their performance within the game. Ranks were assigned so that player's could gauge how close they were to achieving the top score, so this would motivate them to try and learn about the game. The ranks were randomly assigned so that each player were placed 1st, 3rd, and 5th once, to control for the effect of rank on their subsequent behaviour in the game. Participants were then given the choice of seeing either social or asocial information to improve their ship's score (Phase II). The social and asocial information was described as either risky or safe (see below). Once participants had seen this information, they could then construct another ship (Phase III) using items collected via social or asocial information during Phase II. Once participants had finished Phase III, they were told that their score had been saved and their best score out of the three Rounds would be revealed at the end of the experiment. Participants then entered Round 2, where they went through Phases I, II and III again. Participants played three Rounds altogether to ensure that each player was assigned each rank once (1st, 3rd and 5th).

Phase I

Spaceship Construction:

The spaceship was constructed by selecting tiles from a grid of 30 available items, arranged into themes of "Crew, Cargo, Engines, Shields and Lasers" (See **Fig.2.1**). Players had two minutes to explore these items and choose ten to place on to the spaceship template on the right. The only "rule" was to choose at least one crew member and one engine. Players were given no information on how to achieve a good score for their spaceship or on which items were best.

***Figure 2.1**: Participants' view of their first building attempt of the experiment (Phase I)*

Rank Assignment:

After finishing their first spaceship, each player was given a spaceship score and "rank" in the league table (see **Fig.2.2**). Ranks were assigned to control for players' confidence in their ship and to ensure they were motivated to attend to information to help them to achieve the top score. To control for how rank affected subsequent choices in the game, each player was randomly assigned 1st, 3rd and 5th rank once in the game; a different rank in each round. Players were not informed that the ranking was randomly assigned and were told that this was based on their spaceship's score.

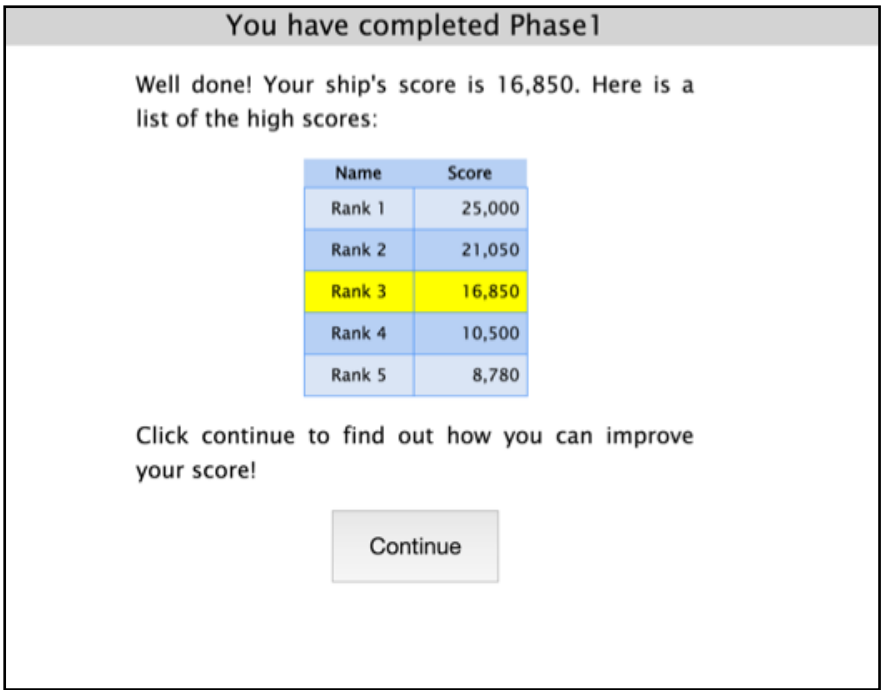


***Figure 2.2**: Participants' view of their rank assignment at the end of Phase I*

Phase II

After seeing their ship's score, players were given the option of either using social or asocial information to help improve their score when building their next spaceship. The social and asocial information was manipulated to seem like a safe or risky choice to the participant, depending on which condition participants were in (see below).

Risky and Safe Manipulations:

A participant was randomly assigned to either Condition 1 (“Social Risky”), Condition 2 (“Asocial Risky”), or the control condition (“Control”) for the duration of the experiment. In Condition 1, social information was “risky” and asocial information was “safe.” Social information (i.e “seeing other players’ ships”) was risky, in that choosing this option would lead to a large increase or decrease in score (see **Fig.2.3**). Asocial information (i.e “spaceship scrapheap”) was safe in that participants’ score was guaranteed to remain the same or increase slightly.

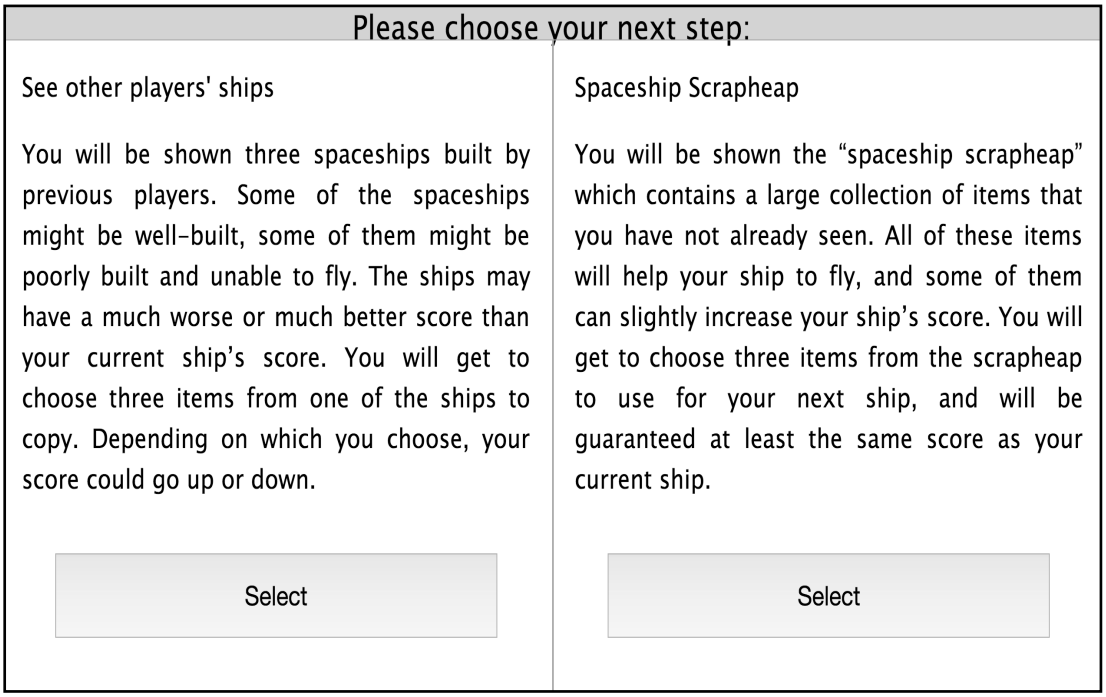


Figure 2.3*: Participants’ view of the social and asocial choice, in the Social Risky condition*

In Condition 2, social information was ‘safe’, and asocial was ‘risky’ (“Asocial Risky”). Social information (i.e “seeing other players’ ships”) was safe, in that choosing this option would guarantee participants’ score to remain the same or increase slightly. Asocial information (i.e. “spaceship scrapheap”) was risky in that choosing this option could lead to a large increase or decrease in score (**Fig.2.4)**.

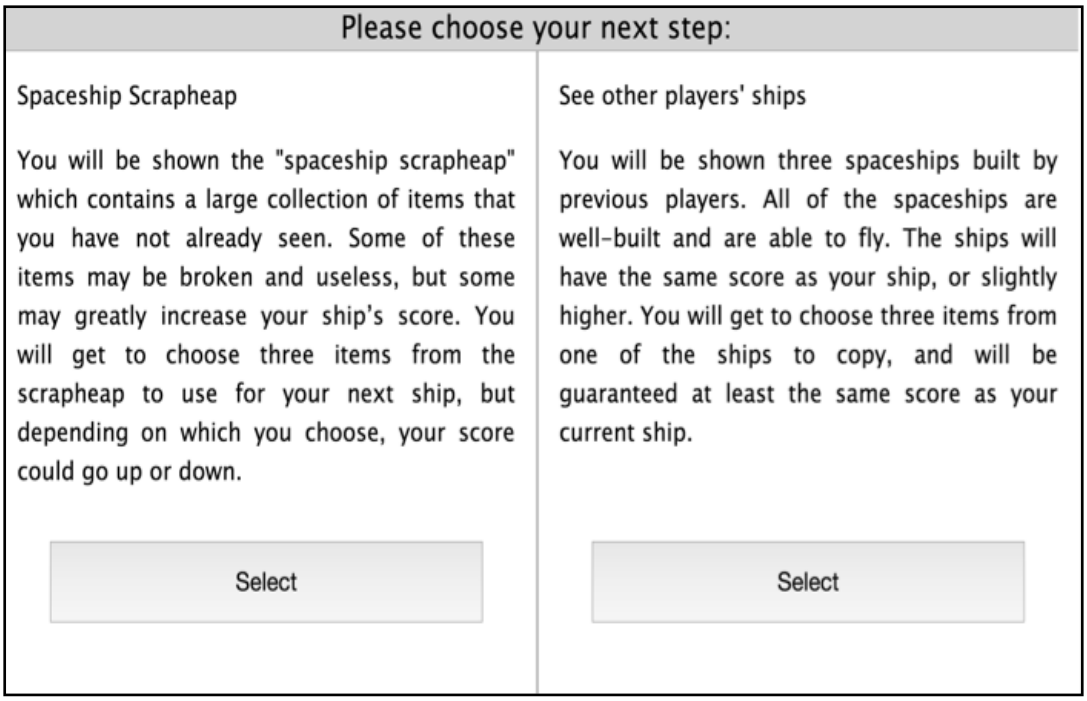


Figure 2.4: *Participants' view of social or asocial choice in the Asocial Risky condition, n.b. order of display is counterbalanced within the task*

In the Control Condition, both options were 'safe,' in that both options guaranteed participants' score to remain the same or increase slightly. This wording reflects the "Bounded Risk Distribution" model in which individuals are expected to maximize their probability of reaching a certain goal while minimizing their probability of falling below a certain threshold (Wang 2002). In this instance, participants are trying to achieve the goal of a top score, and want to minimize their chance of falling below this threshold in order to achieve a bonus payment. Therefore, although the risky options can be interpreted as having a lower expected payoff, this may still be a rational choice if participants are needing to greatly improve their score, due to the potential large increase in score.

Social or Asocial Information Choice:

Social information consisted of viewing three spaceships built by previous players and their associated scores (see **Fig.2.5**). Players could then choose one out of these three spaceships to "copy," and select three items to keep for use in the next building phase.

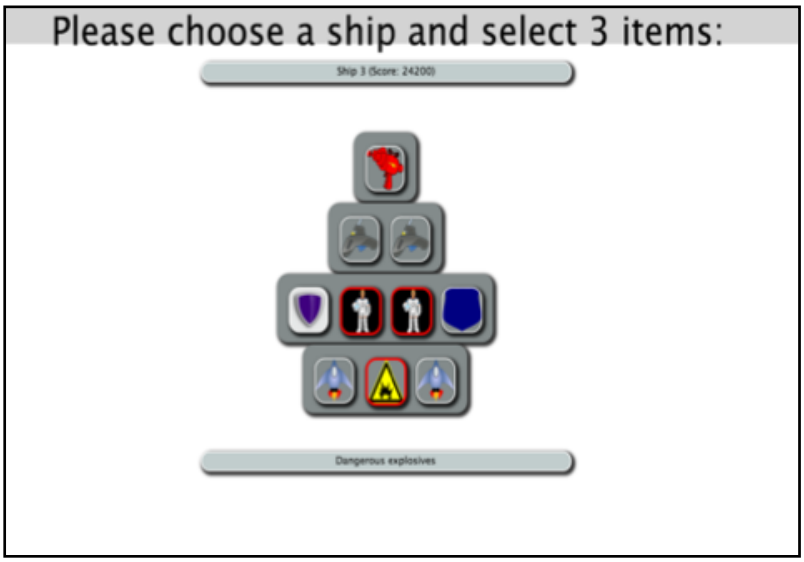


***Figure 2.5**: Participants' view of the social choice, viewing another player's complete ship and score.*

Asocial information consisted of viewing the "spaceship scrapheap" (**see Fig.2.6**) in which a collection of 30 previously unseen items were presented. The participant could reveal up to ten of those items within a time limit of two minutes, and choose three items to keep for use in the next building phase.

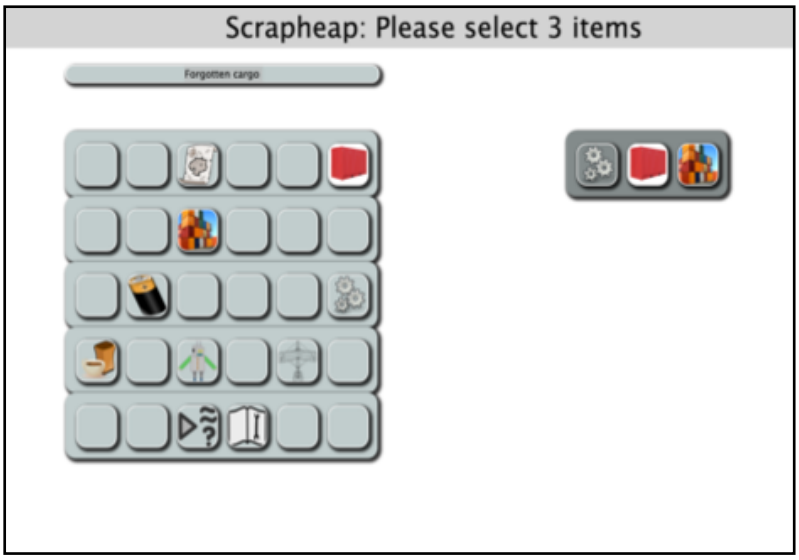


***Figure 2.6**: Participants' view of the asocial option, the "spaceship scrapheap"*

Phase III

After choosing three items from either the spaceship scrapheap (asocial) or from other players' ships (social), participants had to build another ship within a time limit of two minutes, similar to Phase I. The only difference to Phase I was that participants already had the three new items chosen from Phase II placed on their ship, and so only had to select 7 more items from the left (see **Fig.2.7**).

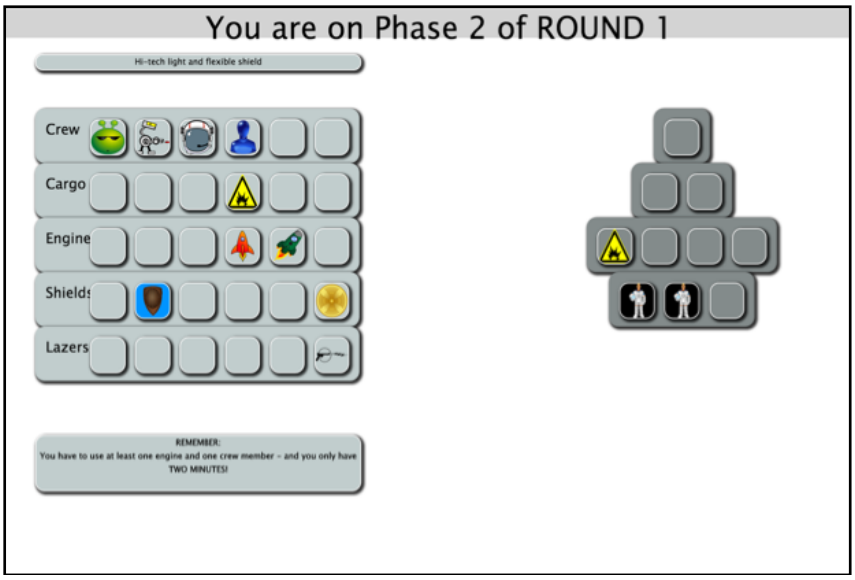


Figure 2.7: Participants' view of Phase III; building another ship using three items from their previous social or asocial choice. n.b. This is described as 'phase 2' for participants due to how the task is explained to the participants.

When participants completed this ship, they were told that their final score had been saved from that round, and their best score from all three rounds would be displayed at the end of the game (see **Fig.2.8**). Participants were not presented with their second score immediately so that participants' did not get any feedback on whether choosing social or asocial learning was a good choice. This ensured that participants' subsequent choices in round 2 and 3 of the game were not influenced by their choice in round 1 or 2.

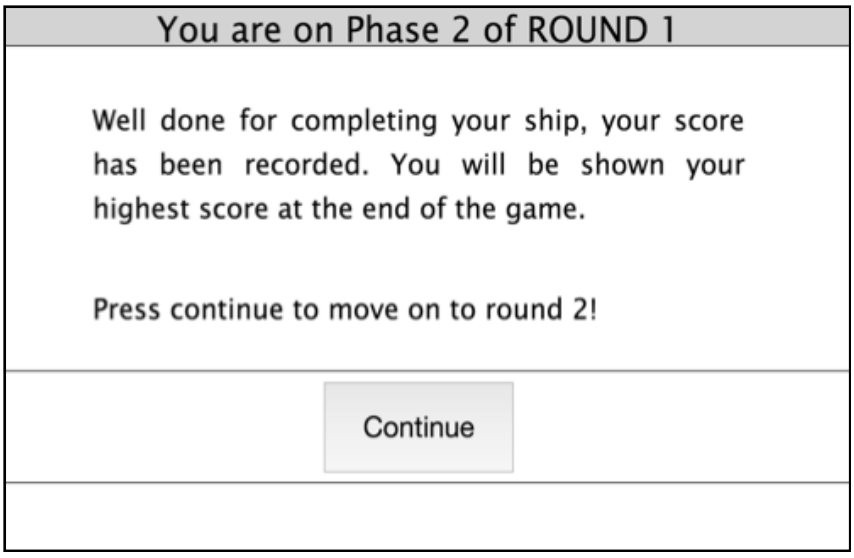


Figure 2.8: Participants are told that their score has been saved and they will see their top score at the end of the game.

Risky-impulsivity Measure:

Participants were given a twelve item risky-impulsivity test (Campbell & Muncer 2009) at the end of the task to test if there was a sex difference in risk-taking in the sample of participants (see **Fig.2.9**). Participants had to rate out of 5 how likely they were to take part in each of the twelve activities.

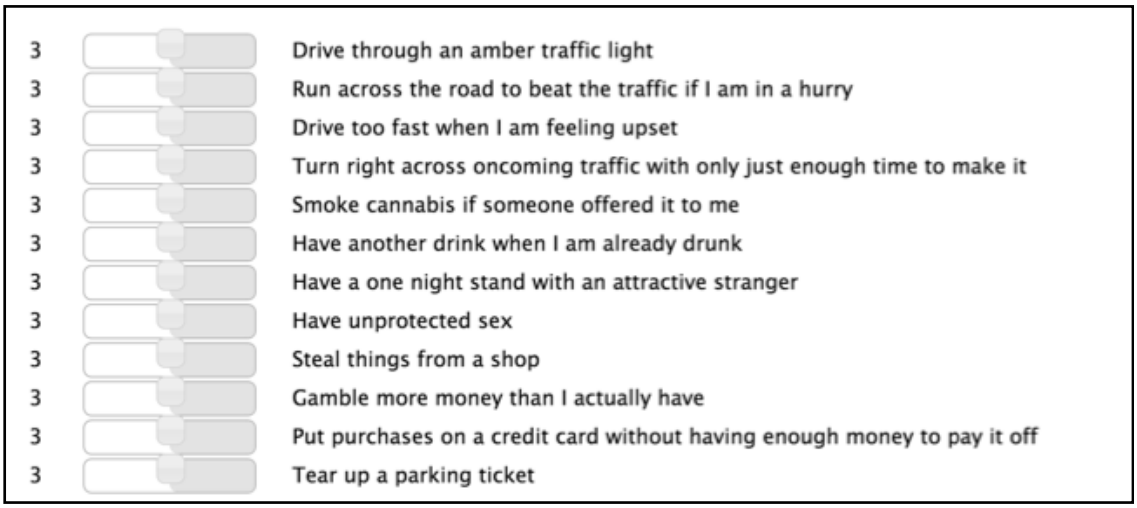


***Figure 2.9**: Twelve risky-impulsivity items as seen by participants.*

2.3 Analysis

All analysis code is available at www.github.com/lottybrand22/GH_RiskySpaceships

The participants' decision to use social or asocial information was modelled using Bayesian binomial logistic regression with the map2stan function in the Rethinking package in R (McElreath 2016). The probability that a participant chose social or asocial information in the task was modelled as a Bernoulli variable (appropriate for binary data, social = 1, and asocial = 0). The full model included an effect for sex (males = 0, females = 0; males are represented as the baseline), an effect for condition one and condition two, an interaction between sex and each condition, and an effect for assigned rank. The control condition was represented in the model as the baseline, so that any effects from condition one or condition two are in relation to males in the control condition.

Model predictions were calculated by averaging across all candidate models weighted according to WAIC weight. Candidate models were chosen based on a priori hypotheses formulated before data collection (see **table 2.1**).

Hypothesis	Model	Parameters included
Null	1	intercept
Full	2	intercept + sex + condition1 + condition2 + sex*condition1 + sex*condition2 + rank + personality
Sex and Condition Interactions predict choice	3	intercept + sex*condition1 + sex*condition2
Sex, and Sex and Condition Interactions predict choice	4	intercept + sex + sex*condition1 + sex*condnition2
Sex and Conditions predict choice	5	intercept + sex + condition1 + condition2
Only Conditions predict choice	6	intercept + condition1 + condition2
Only Sex predicts choice	7	intercept + sex

Table 2.1*: Candidate models and included parameters based on a priori hypotheses*

Although not the primary focus of the study, I wanted to ensure that participants were motivated and attending to the task appropriately (in line with the proposed “Bounded Risk Distribution” model (Wang 2002)). Thus, I ran an additional model with risky/safe choice rather than social/asocial choice as the outcome variable. This model excluded data from the C condition, because both options in this condition were safe and therefore no risky choice could be made. The risky/safe choice was modelled using a Bayesian binomial logistic regression with rank, sex and a sex*rank interaction as predictors. Finally, I modelled participants' risky impulsivity scores using a Bayesian linear model, with sex as a predictor variable, to check whether our sample displayed the expected sex difference in risky impulsivity. To check that the sex difference was not due to the two questions that were explicitly referring to sexual activity, I also conducted the analysis with these questions removed.

2.4 Results

All model estimates are reported with 89% credible intervals (CIs), which are the default in the *Rethinking* package (McElreath 2016). The CIs provide an upper and lower estimate around the mean of the parameter estimate and encompass 89% of the posterior distribution. This method contrasts with the traditional use of 95% confidence intervals in null hypothesis testing. Using 95% intervals would not change the interpretation of our results, because we are using a model comparison approach, and the size of the credible intervals does not affect which models best fit the data. All error bars are 89% credible intervals and can be interpreted as the region within which the model expects to find 89% of responses, given the data and the assumptions in the model.

2.4.1 Risky Impulsivity Model

There was a sex difference in the risky impulsivity test in our sample, in that women scored lower on the risky-impulsivity measure than men **(*β mean* = -0.06, 89% confidence interval: -0.09, -0.04)**. Mean risk-taking score out of 60: males 27.24(7.59), females: 23.41(6.97), cohen's d: 0.52. This small sex difference was still present, yet reduced, when removing the two questions (Q7 & Q8) that specifically relate to sexual activity ***(β mean = -0.04, 89% CI: -0.07, 0.00).*** Mean risk-taking score out of 60: males 22.29(6.56), females: 20.12(6.02), cohen's d: 0.35. A density plot of female and male risky-impulsivity scores on the full 12 questions is shown below (**Fig. 2.10**).

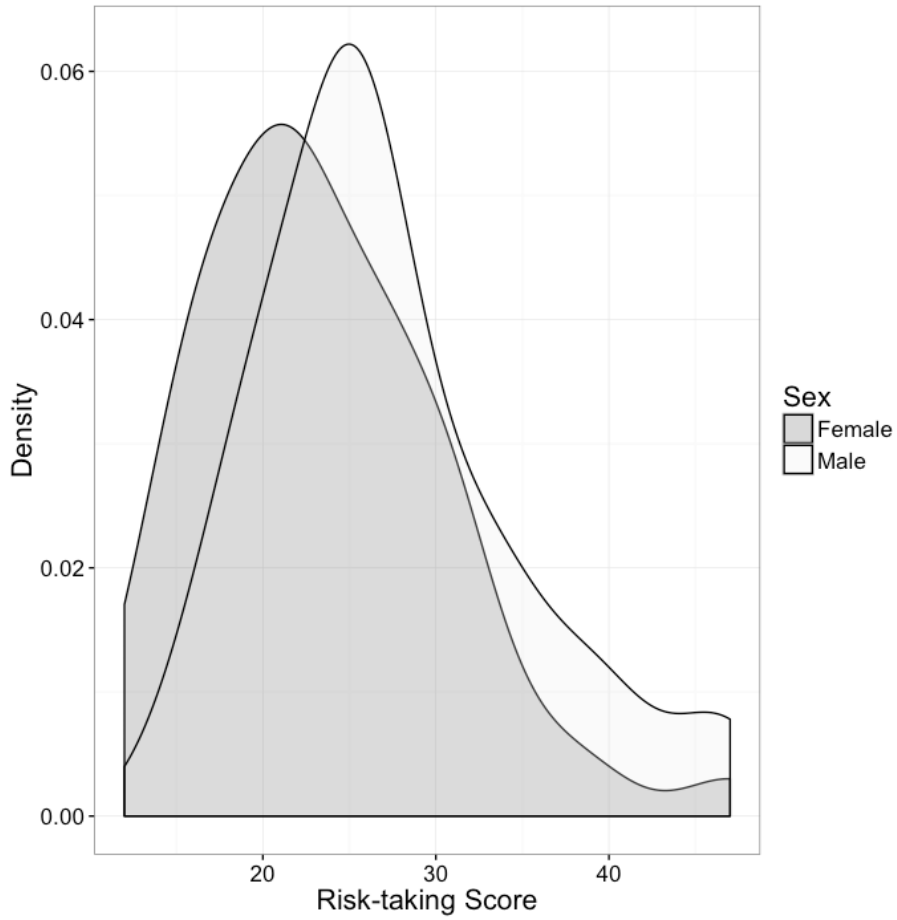


***Figure 2.10**: Density plot showing men and women's risky-impulsivity scores*

2.4.2 Social Learning Model

When modelling the probability of choosing social or asocial information, the model with the lowest WAIC value included an effect for sex and an effect for the interaction between sex and the Asocial Risky condition. A comparison of all model fits is shown below (**table 2.2**) ordered by lowest WAIC score.

Model	Hypothesis	Parameters included	WAIC (±SE)	Akaike weight
1	Null	Intercept	360.3 (5.10)	0.00
2	Full	Intercept + sex + AR + SR + sex*AR + sex*SR + rank + personality	357.1 (10.36)	0.01
3	Sex and condition interactions predict choice	Intercept + sex*AR + sex*SR	351.1 (9.11)	0.25
4	**Sex, and sex and condition interactions, predict choice**	**Intercept + sex + sex*AR + sex*SR**	**349.6 (9.77)**	**0.52**
5	Sex and condition predict choice	Intercept + sex + AR + SR	354 (8.92)	0.06
6	Only condition predicts choice	Intercept + AR + SR	352 (8.77)	0.16
7	Only sex predicts choice	Intercept + sex	362.3 (5.32)	0.00

***Table 2.2**: Models ordered by WAIC value, the lowest value indicates a better fit of the model to the data.*

A table of model coefficients for the best model (Model 4) is shown below (**Table 2.3**) as well as a table of the model coefficients for the full model (Model 2) for comparison (**Table 2.3b**).

Parameter	Mean estimate	StdDev	lower .89	upper.89
Intercept	-0.41	0.19	-0.72	-0.10
Sex	-0.72	0.41	-1.36	-0.05
Sex*AsocialRisky	1.72	0.46	1.03	2.51
Sex*SocialRisky	0.62	0.47	-0.09	1.40

Table 2.3: *Parameter estimates for the best-fitting model, model 4.*

Parameter	Mean estimate	StdDev	lower .89	upper.89
Intercept	-0.36	0.68	-1.46	0.71
Sex	-0.84	0.50	-1.63	-0.05
Sex*AsocialRisky	1.59	0.70	0.46	2.70
Sex*SocialRisky	1.03	0.67	-0.10	2.07
Rank	0.05	0.16	-0.22	0.29
AsocialRisky	0.18	0.52	-0.66	0.98
SocialRisky	-0.42	0.48	-1.19	0.33
PerosnalityScore	-0.09	1.07	-1.85	1.59

Table 2.3b: *Parameter estimates for the best-fitting model, model 4.*

A plot of these coefficients is displayed below **(Fig.2.11)**.

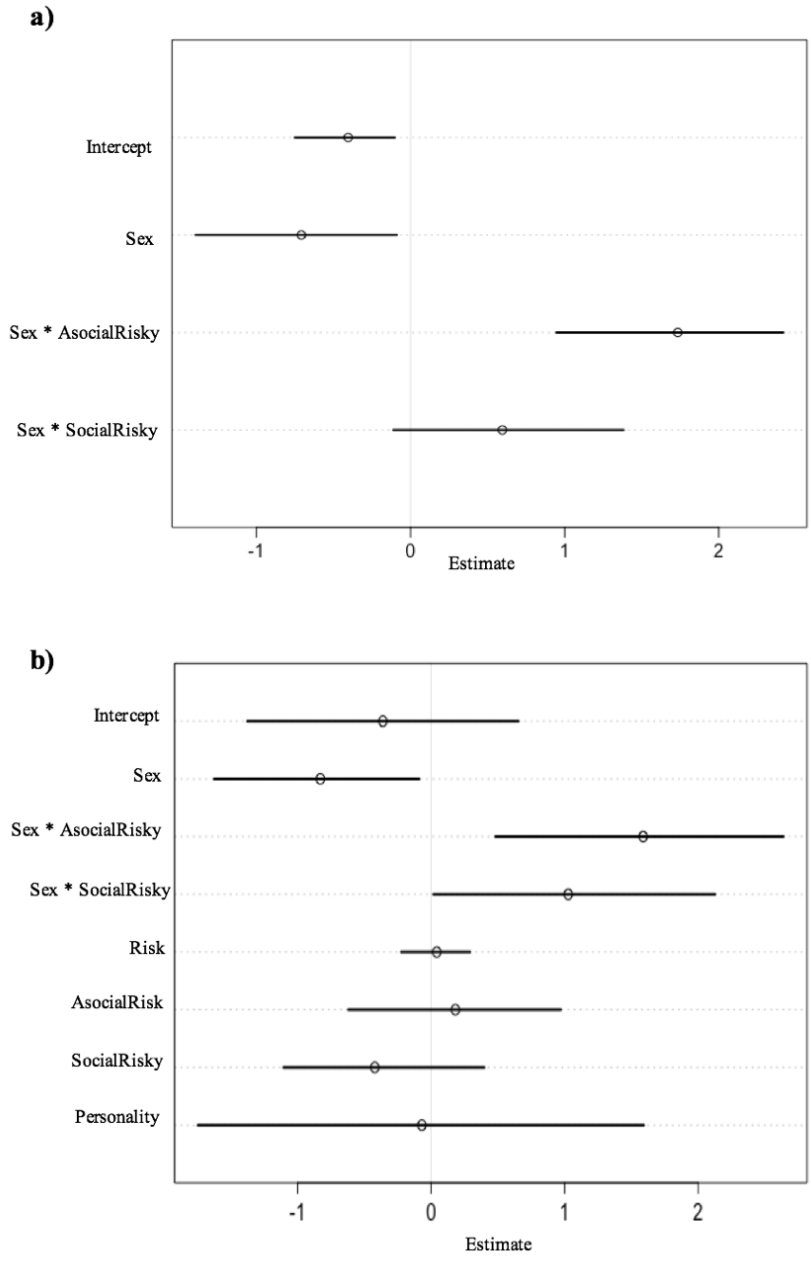


Figure 2.11*: Model coefficients from the best model plotted for each parameter in the model (a) and for the full model (b).*

These model estimates show that the sex and asocial risky interaction had a strong effect in the model **(*β* = 1.76, 89% CI: 1.03, 2.51)**. This effect can be seen more clearly in the model predictions plotted below (**Fig. 2.12)**. Model predictions were generated by averaging across several candidate models, and weighting model predictions according to WAIC values (McElreath 2016).

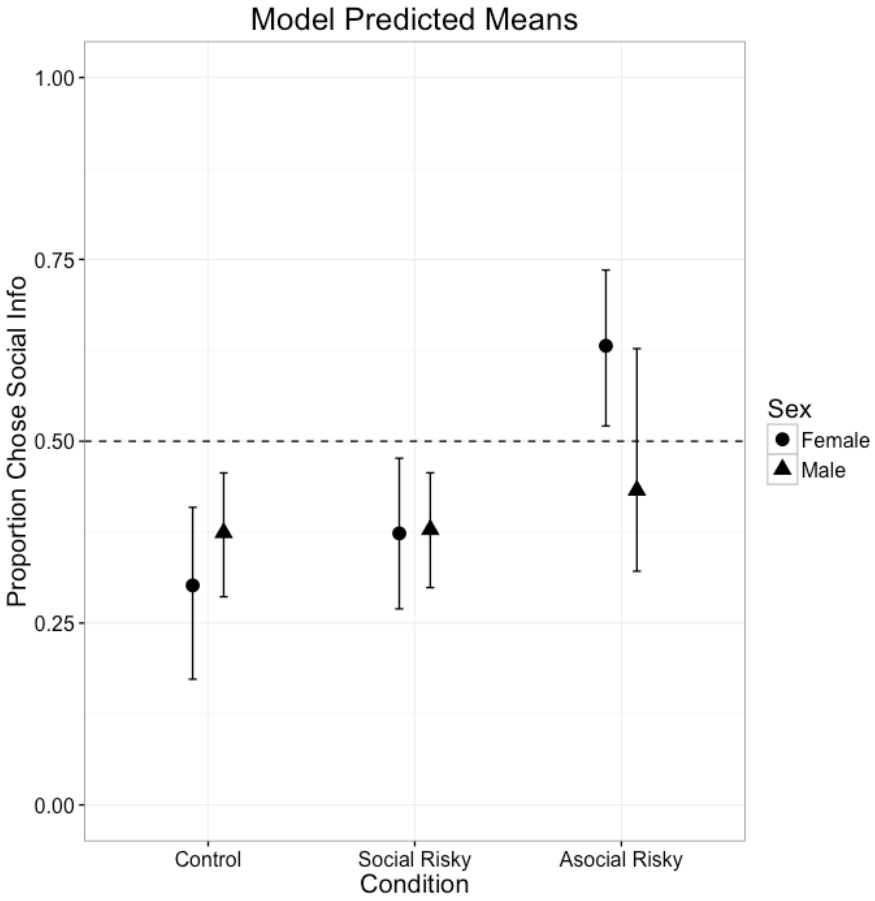


***Figure 2.12**: Predicted mean proportions that chose social information plotted according to condition and sex. Error bars show 89% CI's*

In the control condition, both men and women showed a preference for using asocial information **(men: β = -0.41, 89% CI: -0.72, -0.10; women: β = -0.72, 89% CI: -1.36, -0.05**). In the Social Risky condition, neither men nor women's choice changed compared to the control **(β = 0.62, 89% CI: -0.09, 1.40)**. In the Asocial Risky condition, women are choosing social information more often than asocial, whereas men still show a preference for asocial information **(β = 1.76, 89% CI: 1.03, 2.51).**

2.4.3 Rank Model

In order to check if assigned rank affected participants' choices, and whether rank interacted with sex, I ran a separate model with risky/safe choice as the outcome variable, rather than social/asocial choice. I did not predict rank to affect social or asocial choice, but assumed that it may affect participants' willingness to choose the risky or safe option. This model excludes data from the control condition, as both options in this condition were "safe" and therefore there was no risky choice to be made. The results for this model are shown in **table 2.4** below.

Parameter	Mean estimate	StdDev	lower .89	upper.89
Intercept	-2.38	1.14	-4.20	-0.56
Sex	1.12	1.23	-0.71	3.13
Rank	1.14	0.55	0.22	1.95
Sex*Rank	-0.77	0.69	-1.98	0.19

Table 2.4*: parameter estimates of the rank model*

These results show that across the two conditions in general, participants are choosing the safe option **(*β* = -2.38, 89% CI: -4.20, -0.56)**, and women are not choosing differently than men **(*β* = 1.12, 89% CI: -0.71, 3.13)**. However, when participants are the lowest rank (rank 5) they are more likely to make a risky choice **(*β* = 1.14, 89% CI: 0.22, 1.95)**. There was not strong evidence for an interaction between sex and rank in the model **(*β* = -0.77, 89% CI: -1.98, 0.19)**.

2.5 Discussion

I found that women preferentially chose to use social information when asocial information was risky, whereas men showed a preference for asocial learning even when it was the risky option. Both sexes showed a preference for asocial information in the control condition, and this preference remained when social information was risky too. Rank did not predict participants' choice to use social or asocial information, but it did predict whether they chose the risky or safe option, in that participants who were ranked lowest were more likely to choose the risky option. There was no interaction between sex and rank, showing that women and men were responding to rankings in a similar way. I found a sex difference in a risky-impulsivity measure in that men scored higher on the risky-impulsivity scale on average than women.

When asocial information was risky, women preferentially chose to use social information, whereas women in the control condition showed a preference for asocial information. However, this pattern did not occur in men, as men showed a preference for asocial information across all three conditions. This result fits in with previous findings that women are less risk-taking than men (e.g. Cross, Copping & Campbell 2011). Furthermore it is consistent with the interpretation that sex differences in risky behaviour can be better explained by differences in punishment sensitivity rather than reward sensitivity, as women changed their behaviour to avoid the risky option when it became risky, whereas men did not change their behaviour. Thus this pattern of results may be better explained as women being more sensitive to the possible punishment of a low score in the risky option, rather than men being more sensitive to the possible reward of a large score in the risky option.

Similarly, men's preference for asocial information did not change according to riskiness across the conditions. A likely explanation for this result is that in our experiment, the risk

was salient enough to alter women's behaviour, but not high enough to alter men's behaviour. Given the sex difference in risky-impulsivity in this population, it is reasonable to expect men and women to have different cut-off points at which a choice is perceived to be risky enough to alter one's behaviour. Future studies could implement a more complex pay-off structure, so that the point at which both women and men change their behaviour according to riskiness could be observed and compared. Alternatively a more continuous, sensitive measure of how participants are perceiving risk in the task could be implemented, such as that used in a study of overconfidence by Muthukrishna and colleagues (in press) and a recent delay-discounting study (Yim et al. 2016). A similar method could be used to assess participants' estimation of risk within a task. More accurate and continuous measures of risk perception could allow us to discover different within-sex cut-off points for when participants change their behaviour.

Participants also had a preference for asocial information in the control condition. This preference fits with previous experimental findings showing that social information is often underused even when adaptive (Mesoudi 2011; Morgan et al. 2012). Although theoretical models have shown that social learning is the most adaptive strategy over many generations of simulated evolution (Rendell et al. 2010), in our study and others it appears participants are keen to solve the task for themselves initially. Furthermore, in our task it could be the case that the asocial option is perceived as inherently more appealing, novel or interesting, and thus the majority of the participants chose to learn asocially overall. The fact that participants show a preference for the asocial option in the control condition could also help to explain why participants had an asocial preference in the Social Risky condition too, as there is even more incentive to choose the asocial option when it is also the safe option.

When social information was risky, neither men nor women's behaviour changed compared to the control condition. One reason for this could be because the social information was not

perceived as very risky, even though it was described as such. Although participants knew that their score could increase or decrease dramatically by choosing the social information, it could be perceived as inherently "safe" as the social information has presumably been derived by other participants. Although they are told that other participants' ships could have very high or very low scores, the fact that they are viewing information from other individuals could be perceived as a helpful strategy regardless of how variable this information might be. This interpretation of social information being inherently "safe" or beneficial ties in with previous social learning research that shows social learning to be the most adaptive strategy in a social learning strategies tournament (Rendell et al. 2010). Learning from others in a new environment that you have not yet learned about is usually the safe, adaptive strategy (Henrich & Boyd 1998; Rendell et al. 2010).

If it is the case that the asocial option is more appealing in our task, the fact that women, but not men, forfeited the more appealing option when it was made riskier is still of interest. However, men were not more sensitive to the more rewarding choice than women, as both men and women preferred the perhaps more appealing, novel, or enjoyable asocial option in both the control and the Social Risky condition. It could be the case that the social option is perceived as the more boring option in this task, supporting previous findings of sex differences in choosing the "safe but boring" option compared to a more novel, yet potentially more dangerous, option (e.g. Tellegen 1982). This interpretation could be tested in future studies by trying to design a task in which the social choice is the most inherently appealing when controlling for risk, or by controlling for novelty value in various task conditions.

Participants with the lowest rank were more likely to choose the risky option, regardless of sex. This is consistent with results from research in economic games in which participants are more likely to take a risk if they are performing poorly in comparison to other

participants, as they have less to lose and more to gain (Mishra, Him & Lalumiere 2015). However, low ranking individuals were not more likely to choose to learn socially over asocially. Although copy-when-unsuccessful is an adaptive social learning strategy (Hoppitt & Laland 2013), another social learning experiment also found that participants were no more likely to copy when given a lower rank (Toelch et al. 2011). Furthermore, there was no guarantee in our task that participants would see a ship from a high ranking individual, even if they did choose to copy. The risky-impulsivity measure scores also did not predict use of social or asocial, or risky or safe sources of information. However, individual differences are more likely to occur in larger samples, and could be more apparent in a more relevant risk-taking measure such as one specific to economic risk.

Another avenue for future research could examine the link between perceived rank and risk-taking more closely by providing a more continuous and varied rank assignment. Our assignment of rank was the most basic in that participants could either be top, bottom, or middle rank (1,3,5). The relationship between confidence and risk-taking could be explored by measuring how perceived rank affects confidence, which in turn may lead to riskier or safer choices within a task. As confidence has previously been found to affect social information use, exploring how manipulating rank affects confidence would be of interest. In support of previous work, the sample of participants in this study showed the predicted sex difference in risky-impulsivity in that men had higher risky-impulsivity scores on average than women (Campbell & Muncer 2009). However, it is worth noting that this small sex difference was reduced when the two questions explicitly related to sexual activity were removed from the 12 items. Although the 12-item scale that we used has been previously validated as a 'risky-impulsivity' measure, the confounding questions on sexual activity, which are inherently riskier to men than women anyway, may deem this scale inappropriate for our type of study. Thus in future studies, a scale specific to 'economic risk-taking' (ie behaviours related to gambling and personal finances) may be more appropriate.

Chapter 3:
Confidence and Conformity

3. Confidence & Conformity

Introduction

The aim of these studies was to directly test whether men and women's confidence levels on a mental rotation task can be altered using stereotype threat, and to see if this difference in confidence then leads to a difference in conformity levels. A recent study found that women's confidence was lower than men's, and consequently conformity was higher, in a mental rotation task but not in a letter transformation task (Cross et al. 2016). This presence of a sex difference in one of the tasks but not the other indicated that stereotype threat may have affected women's confidence in the mental rotation task. I used a different version of a mental rotation task (used in one previously published study; Alexander & Evardone 2008), which uses 3D images of human figures, rather than 3D shapes made out of blocks. Alexander and Evardone (2008) found a sex difference in the version which used images of human figures, with males performing better than females on average, but the difference was not as marked as the classic block version. Thus, I used the human figure version so that expectations about how men and women perform on the task could be manipulated, as it is a less well-known version than the standard blocks design.

In a similar vein to Sharps, Price and Williams (1994), who described a drawing task as either relying on geometrical skills or memorising ability, I attempted to manipulate expectations about the task by describing it as either a spatial-shifting task or a perspective-taking task. The descriptions are based on commonly held stereotypes that are either positive for men or women, in that men are believed to be better at spatial tasks (Hirnstein, Andrews & Hausmann 2014), and women are believed to be better at perspective-taking tasks (Mohr, Rowe & Blanke 2010; Davies & Stone 1995). These descriptions are combined with explicit statements about men and women's previous performance on this task (i.e women have been found to perform better than men, or men have been found to perform

better than women). A previous meta-analysis found that subtle primes are more effective at altering performance than explicit ones (Nguyen & Ryan 2008); thus we employed distractor material about the task alongside the primes (i.e. reporting the University at which the bodies version originally took took place), and embedded them within the instructions so as not to be too conspicuous to the participants.

Participants were required to state whether two images were rotations of the same or different body positions, and rate their confidence in their answer. I predicted that women's average confidence levels would be higher in the condition in which they were primed as 'females better' compared to women's average confidence levels in the control group (no prime). Similarly I predicted that women's average confidence levels would be lower in the 'males better' condition compared to women's average confidence levels in the control condition. I expected the opposite pattern for males. I predicted participants' confidence to be correlated with their performance (i.e. whether they got the answer correct or incorrect), as their confidence ratings should be a rough estimate of how they are performing. I also included a manipulation check questionnaire at the end of the study to get an indication of whether participants remembered, believed, and understood the stereotype threat primes that were embedded in the task instructions.

After successfully manipulating participants' confidence in a novel mental rotation task using stereotype threat in Study 1, I then used these same stereotype threat primes in an identical task with an added social information component (Study 2). After giving their answer and confidence rating, participants were shown other peoples' answers for each trial, and given an opportunity to change their answer to conform to the majority's answer. Social information was always displayed as a majority of people (out of 12 possible people) giving their answer "yes" or "no" to that trial. Participants were then reminded of their own original answer, and asked again if they thought the images matched or not. "Switching" was defined as when a participant changed their original answer after seeing the social information. Social

information was displayed randomly, thus not all majorities disagreed with the participant's original answer. The social information aspect of the task was based on Cross and colleagues' (2016) methodology. I predicted that lower confidence ratings would lead to higher rates of conformity, as was the case in two previous studies with similar methodology (Morgan et al. 2012; Cross et al. 2016).

The study by Morgan and colleagues (2012) found that lower confidence predicted higher conformity on a mental rotation task, indicating that people were using a "copy-when-uncertain" social learning strategy. Cross and colleagues (2016) then found that although sex had no direct effect on switching, it did have an indirect effect on switching mediated by confidence (see **Fig.3.1**). This mediating effect of confidence was not found on the letter transformation task. Therefore I wanted to examine whether confidence was also a mediating factor in this version of the mental rotation task, in which human figures rather than blocks were used.

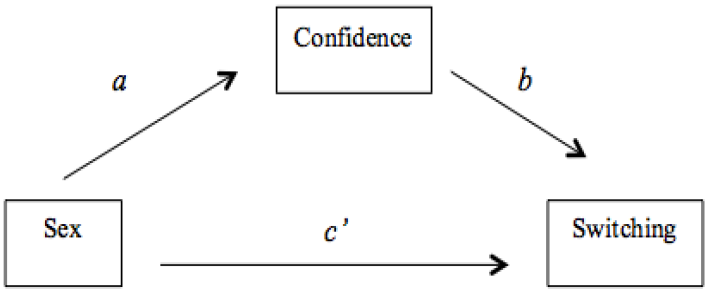


Figure 3.1*: Diagram of mediation analysis in which the effect of sex on switching (c') is mediated by confidence. Path c' represents a product of two paths, i.e. path 'a' and path 'b'.*

In Study 3, I repeated Study 2 but with an online sample rather than administering the task in person with participants recruited from the University. This was to check that the different results in Study 2 to Study 1 were not due to a difference in sample. Thus, Study 1 was aimed at manipulating participants' confidence. Study 2 was aimed at examining the effect of confidence in conformity, and Study 3 was a replication of Study 2 using an online sample.

Study 1: Manipulating Confidence

In Study 1, I tested whether the stereotype threat primes would have an effect on participants' confidence ratings compared to the control condition with no primes. Specifically, I predicted that women in the 'males better' condition would have lower confidence than women in the control, or than women in the 'females better', condition. Although, as previously mentioned, results on how men react to stereotype threat are mixed, I predicted the opposite pattern for males, based on results from studies with comparable stereotype threat primes to the ones used in our study.

3.1 Methods

3.1.1 Participants

120 participants (60 males and 60 females, 20 of each sex in each condition) were recruited through Qualtrics, an online survey software and sample recruitment service. Participants were randomly assigned to each condition, as it was a between-subject design. A sample of participants aged over 16 from the UK was pre-specified with Qualtrics, and all participants had to read an information page and give their full consent before taking part in the experiment online. Participants knew that they could withdraw from the study at any point without explanation. The experimental protocol was approved by the departmental ethics committee (see Appendix for letter of approval). 64% of participants fell into the 31-60 years age category, 26% in the 16-30 range, 9% in the 61-80 category and 1 participant preferred not to disclose their age category. Before the experiment began, participants were asked to provide their age, gender and country of birth. Participants were rewarded £3 for their time and had the opportunity of achieving a further £2 bonus based on performance in the task.

3.1.2 Materials

The stimuli were obtained from Gerianne Alexander, Texas A&M University. Images of human figures had been put into various configurations and rotated using the classic Vandenberg & Kuse (1978) stimuli as a template; rotations of the human figures also included occlusions of some features just as the original Vandenberg and Kuse stimuli did. My presentation of the stimuli differed to that of Alexander and Evardone in that our participants were presented with two images on the screen at once that either matched or did not match; this set-up is most similar to the Shepard and Metzler (1971) style of mental rotation task. I generated 24 stimuli that were made up of 12 matching and 12 non-matching rotations (half of the stimuli are female figures, the other half are male figures) all of which were displayed wearing identical t-shirt and trousers. The stimuli were randomly ordered for each participant.

3.1.3 Primes

All participants were first shown an example of the traditional mental rotation stimuli (see below) and were given instructions on how to complete a classic mental rotation task. All participants were then shown an example of the bodies stimuli. Primes were embedded in the instructions, as shown in the following excerpts stating the exact wording seen on screen by the participants. The sections in red are wording that changed in the two conditions, and were absent in the control condition. The example wording is for Condition One (‘Females Better’):

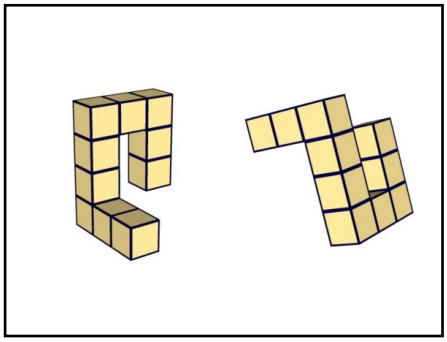
Condition one ('Females better'):

Here is an example of a widely used spatial task in which participants have to decide if the second image is a rotated version of the first image, or if it is a ***different*** *object altogether.*

This task relies on spatial shifting skills, and men have been found to perform better than women.

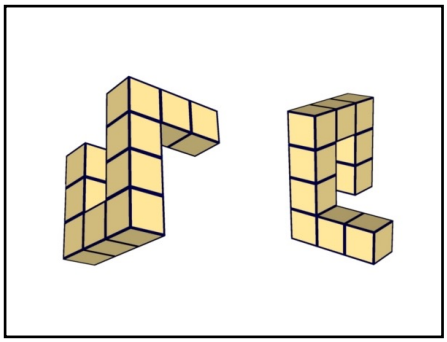
Examples:

Do these images match?



These images match. They are the same ***except one is rotated.*** *Therefore the correct answer is* ***YES.***

Do these images match?



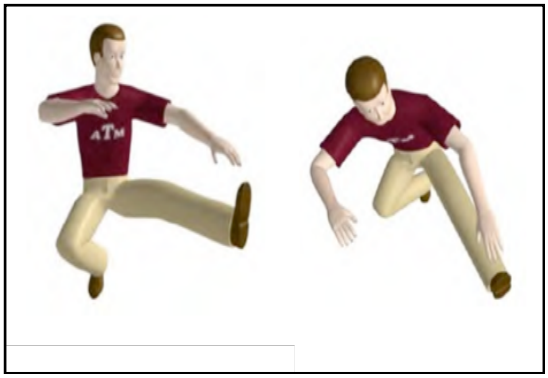
These images do not match. They are two ***different*** *images. Therefore the correct answer is* ***NO.***

Here is a different version of the task that you will see today, in which human figures are used. This new task was developed in 2007 at the Texas A&M University. Participants still have to decide if the two images are the same or different.

Unlike the previous task, this version relies on your ability to take on another's perspective, and women have been found to perform better than men.

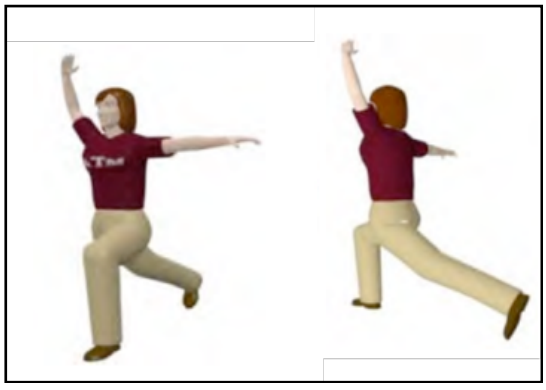
Examples:

Do these images match?



*These images match, they are the same **except one is rotated.** Therefore the correct answer is **YES.***

Do these images match?



*These images do not match, they are two **different** images. Therefore the correct answer is **NO.***

In Condition 2 ('males better'), the red writing was altered to be the following:

> *"Here is a similar version of the task that you will see today, in which human figures are used. This new task was developed in 2007 at the Texas A&M University. Participants still have to decide if the two images are the same or different.*
>
> *Just like the previous task, this version relies on spatial shifting skills, and men have been found to perform better than women."*

In the control condition, the first red sentence mentioning gender was eliminated, as was the final sentence; no mention of gender or skill-type was made. All participants were told that the information had been manipulated at the end of the experiment during the debrief.

3.1.4 Procedure

After giving consent and providing their age, gender, country of birth and education level, participants then read the instructions as shown above, depending on which condition they had been assigned to. In each trial, participants saw a pair of 3D images of human figures on a computer screen online, as shown above. The participants had to judge whether the two images were identical except for one having been rotated, or whether the two figures were different altogether. The participants saw the pair of images for 4 seconds before being asked "Did these images match?" and responded by clicking either "yes" or "no" on the computer screen. Each participant completed 24 trials. After each "yes" or "no" response, participants were asked "how confident are you that you got that right?" and responded using a 7 point Likert scale (see **Fig.3.2**). Participants were told that rating their confidence as 0 on the scale meant that they guessed their answer, and felt there was a 50% chance that they got the question right. They were told that rating their confidence as 6 on the scale meant that they were 100% sure that they got the question right. The scale pointer began at 3 (mid-way) and participants had to click on the pointer before being able to proceed to the next page. Participants did not get feedback on their performance at any point during the

experiment, so their confidence ratings should be a direct reflection of their estimation of their performance throughout the study.

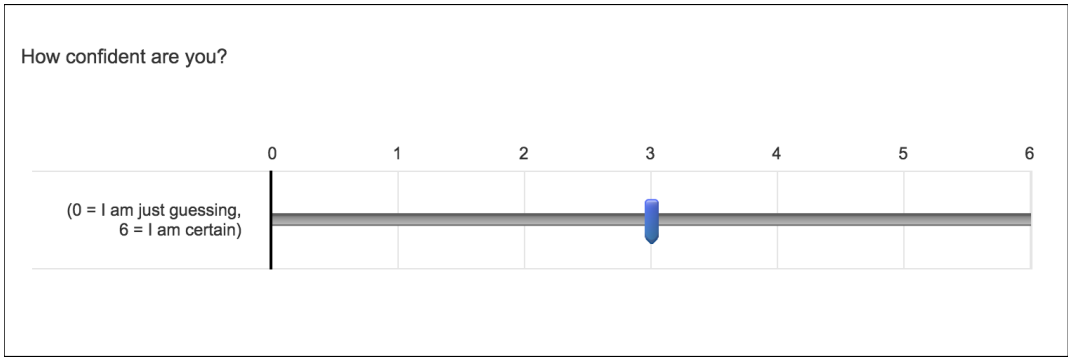


***Figure 3.2**: Confidence scale used by participants after giving their initial answer on each trial.*

3.1.5 Manipulation Check

After the experiment, participants were asked how familiar they were with the stereotype that men perform better at spatial tasks than women previous to taking part in the study. They were also asked whether they believed the stereotype (on a 7 point likert scale). We then reminded them of the manipulation that was within the instructions at the beginning of the experiment and asked them if they remembered reading the information, and whether they believed the information. We then asked them how well they expected to perform on the task compared to members of the opposite and same sex.

3.1.6 Statistical Analysis

The data were analysed using Bayesian generalized linear mixed models (GLMMs), modelling the performance of the participants on the task (ie whether they got each trial correct or incorrect), and their confidence per trial, in R and JAGS using the R2Jags package.

Mediation Analysis

Throughout this chapter I will be using mediation analysis. Mediation analysis is a type of path analysis, akin to structural equation modelling (SEM), in which a third variable is added to the interpretation of a relationship between an independent and a dependent variable. Structural equation modelling (SEM) is a powerful means of comparing different theoretical relationships between variables by comparing how well the different models fit the data. In contrast, mediation analysis is concerned with one particular aspect of a relationship between two variables. Specifically, a mediating variable is intermediate in the causal path between the independent and dependent variable, i.e. there is a causal relationship between the independent variable, the mediating variable, and the dependent variable (MacKinnon 2008). Thus, in contrast to a covariate, a confound, or a moderator, a mediating variable changes the relationship between the independent variable and the dependent variable such that it adds an extra causal step between the independent and dependent variable, via the mediating variable (MacKinnon 2008; MacKinnon, Fairchild & Fritz 2007). Mediation analysis is very common in social psychological research as it is intrinsic to the theoretical frameworks that researchers use, i.e. in understanding how psychological or social factors play a mediating role in affecting behavioural outcomes (MacKinnon, Fairchild & Fritz 2007). Indeed, 59% of articles published in the *Journal of Personality and Social Psychology* from 2005-2009 contained at least one mediation analysis, as did 65% of articles in *Personality and Social Psychology Bulletin*.

Traditionally, the majority of mediation analyses involve following a “causal steps” approach outlined originally by Baron and Kenny (1986), in which the relationship between two variables is tested prior to introducing a potential mediating variable into the analysis. Testing the relationship before and after adding a mediating variable ensures that there is indeed a relationship to mediate (a “direct effect”), and that the mediating variable accounts for part (or all) of the effect (an “indirect effect”). However, many researchers have called for a move away from the emphasis on obtaining significant direct effects, arguing that focusing on the presence of direct effects can lead to misleading or false conclusions (Rucker et al. 2011), as well as preventing researchers from exploring important and interesting theoretical questions (Zhao, Lynch & Chen 2010). Thus, many researchers argue that the only requirement for mediation is that an indirect effect be present (Zhao, Lunch & Chen 2010). Furthermore many researchers have heavily criticized the traditional Baron & Kenny approach (Hayes 2009; Zhao, Lunch & Chen 2010; MacKinnon, Krull & Lockwood 2000; Shrout & Bolger 2002; Rucker et al. 2011) and have demonstrated that it is not only underpowered (Fritz & MacKinnon 2007) but also the least effective means of testing intervening variables (MacKinnon et al. 2002; Hayes 2009).

Examples of how an indirect effect may be present whilst the direct effect is absent include suppression effects, asymmetries in statistical power between the effects, and differing strengths of effects (Rucker et al. 2011). Suppression effects occur when an indirect effect has a sign which is opposite to the total (direct) effect (MacKinnon 2000). For example, the presence of lifeguards at a beach may have a direct negative effect on number of fatalities at the beach (i.e. reducing fatalities). However, there could be a mediating positive effect, in that the presence of lifeguards reduces the perceived risk of swimming in the sea. As the perceived risk of swimming in the sea decreases, the number of people choosing to swim in the sea increases, thus the number of fatalities increases. If the strength of the indirect effect

is large enough, and the strength of the direct effect is small enough, this could lead to a non-significant direct effect whilst a significant indirect effect is present. Furthermore, it could be the case that both effects are present, perhaps the perceived risk only has a small effect, and the overall "net gain" of lifeguard presence is positive, in which case both the indirect and direct effect are "significant," with a small "partial" indirect effect. Nevertheless, this 'partial' indirect effect is still important and worthy of study, as it could have important implications for understanding that number of lifeguards may increase number of swimmers. Accordingly, if there is theoretical evidence or support for the presence of an indirect effect, this effect should be investigated, regardless of the size or presence of a direct effect.

In this chapter, I am basing my research on previous findings that suggest confidence may be a mediating factor, not only in the effect of gender on conformity, but also the effect of stereotype threat on performance. Furthermore, previous work by Cross and colleagues (2016), to which I used similar methodology, found evidence of an indirect effect of sex on conformity, mediated by confidence, but no evidence of a direct effect. Thus, I investigated the mediating variable of confidence on conformity, regardless of the presence of a direct effect of sex on conformity or not. Specifically, I conducted a Bayesian mediation analysis (Yuann & Mackinnon 2009) to assess the indirect effect of stereotype threat on performance via confidence, as well as the indirect effect of sex on conformity via confidence.

Binomial logistic regression vs. Ordered logit

Throughout this chapter, confidence was measured via a 7-point Likert scale (0-6). We chose to model this ordinal data using a binomial logistic regression, with 6 trials. This was in contrast to using an ordered logit model, although both methods were used and compared on pilot data to ensure that they would not lead to meaningful differences in inference. The binomial model with 6 trials can be interpreted as a kind of multinomial model, although each response is treated as one event, rather than 7 separate events. A binomial model with 6

trials can be understood by imagining that each confidence level corresponds to flipping a coin 6 times; heads = “confident” and tails = “not confident.” Thus, a confidence rating of “3” represents landing on “heads” three out of six times. Although seemingly counterintuitive, representing the confidence scale in this way effectively makes the middle of the scale the most likely (i.e. 3/6), and the extremes of the scale least likely (i.e. 0/6, 6/6), which we feel is representative of the scale, particularly as the default position for the slider-scale is 3, and participants have to actively move the slider-scale up and down to select their level of confidence (discussed further in the methods section).

In contrast, modelling the data as their ordered categories requires using a log-cumulative-odds scale with a cumulative link function (McElreath 2016). The cumulative probability of a value is the probability of that value, or any smaller value. In our confidence scale, the probability of a confidence ratings of 3 is the sum of the probability of confidence ratings 1, 2 and 3. Due to this summing of probabilities, our scale cannot include a zero, and thus the scale was shifted from 0-6 to 1-7 to allow for ordinal logit modelling. Using cumulative probability enables the use of a linear model with predictor values whilst ensuring the correct ordering of the outcomes. This requires defining intercepts that represent the cumulative probability of each confidence level. Thus the model output provides 6 estimates of the relative frequencies of each confidence level (when the predictors are set to zero), as well as the estimates for the predictors in the model. Therefore, interpreting the predictor estimates relies on interpreting the relative change in cumulative log-odds for every value of the response variable (McElreath 2016). For this reason, alongside the significant increase in running-time for ordered logit models, we decided to use the binomial model for all subsequent analyses. There was no meaningful difference in inference between the models, and both types of models gave very similar model fits. Examples of both types of model code are openly available at www.github.com/lottybrand22/GH_ConfidenceConformity.

The model of participants' confidence ratings included a baseline value, an effect for sex, condition one, condition two, an effect for performance, an interaction between sex and each condition, and two random effects to allow for variation between individuals and between question numbers. Again, the control condition was represented in the model as the baseline, so that any effects from condition one and two are in comparison to the control condition.

The probability that a participant's answer was correct was modelled as a Bernoulli variable (appropriate for binary data, correct = 1, and incorrect = 0). A binomial logistic regression model was written in JAGS, and ran in R using R2Jags. The model included a baseline value, an effect for sex, condition one and two, interaction between sex and condition, and two random effects to allow for variation between individuals and between trials. The control condition was represented in the model as the baseline, so that condition one and condition two are compared to the control condition separately, and not to each other.

Participants' responses to the manipulation check questions were modelled in the same manner. Whether the participant remembered and believed the task manipulation was modelled as a Bernoulli variable. The 0-6 likert response for how participant's expected to perform compared to the opposite sex was modelled as a binomial with 6 trials in the same manner as the confidence analysis.

All models were chosen a priori based on hypotheses formed before the data were collected. All models ran 3 chains with 61,000 iterations, a burnin period of 1000 and went through thinning of 20. Full model code can be found at www.github.com/lottybrand22 and examples of model output such as posterior distributions of parameter estimates and trace plots of Markov chains can be found in the appendix to this chapter.

All plots are a display of the raw data only, with error bars showing 95% confidence intervals. All results in brackets are parameter estimates from the model output, including the median

and 95% posterior distribution parameter estimates. 95% credible intervals that include zero suggest there is little evidence that the parameter had an effect on the outcome variable. 95% credible intervals that do not include zero suggest that the parameter had a negative or positive effect on the outcome variable in the model, depending on the direction of the estimates.

3.1 Results

3.1.7 Confidence

Sex did not predict confidence in the control condition **(*β median* = -0.47, 95% Credible Intervals: -1.00, 0.08),** but it did predict confidence across all three conditions, in that men gave higher confidence ratings on average than women overall **(*β* = -0.47, 95% CI: -0.77, -0.17, Fig.3.3).** Both men and women gave higher confidence ratings in the 'females better' condition compared to men and women in the control condition **(*β* = 0.43, 95% CI: 0.06, 0.81, Fig.3.3**). There was no interaction between sex and either condition **(condition1: *β* = 0.02, 95% CI: -0.74, 0.75, condition 2: *β* = -0.12, 95% CI: -0.89, 0.62)**. Performance predicted confidence, in that participants gave higher confidence ratings when they answered a question correctly compared to when they answered a question incorrectly **(*β* = 0.27, 95% CI: 0.19, 0.34, see Fig.3.4**)

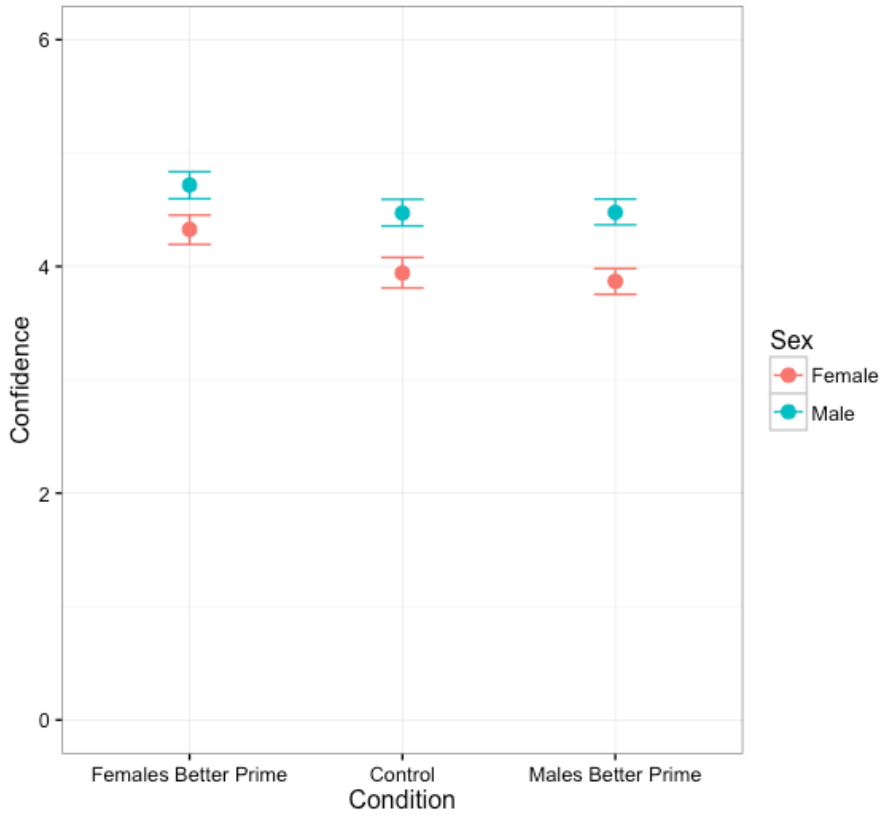


***Figure 3.3**: Plot showing men and women's mean confidence ratings per condition. Error bars represent 95% confidence intervals bootstrapped for non-normal data.*

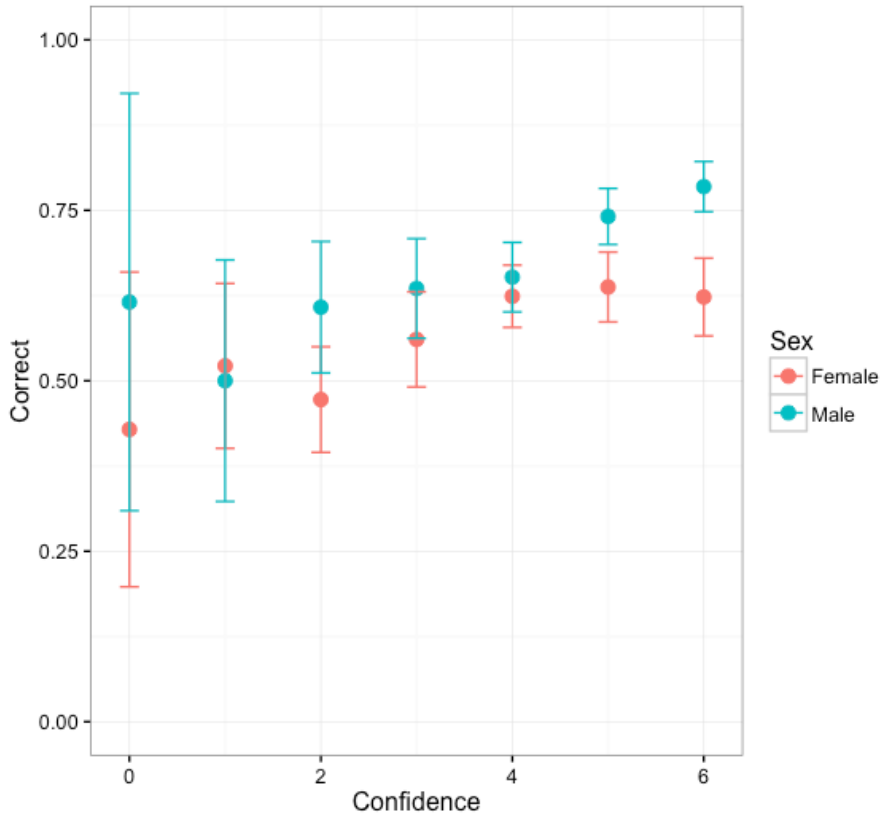


***Figure 3.4**: Plot showing the relationship between men and women's confidence ratings and their proportion of correct trials. Error bars show 95% confidence intervals.*

3.1.8 Performance

Sex predicted performance in the control condition in that men performed better than women (***β* = -0.78, 95% CI: -1.42, -0.15, see Fig. 3.5**) . Neither condition affected performance compared to the control condition, in that participants performed similarly across all three conditions (**Condition 1: *β* = -0.23, 95% CI: -0.68, 0.22, Condition 2: *β* = -0.18, 95% CI: -0.63, 0.28**). There was no evidence of an interaction between sex and either condition, in that both men and women's responses remained similar to the control responses in both conditions (**Condition 1: *β* = 0.18, 95% CI: -0.68, 1.06, Condition 2: *β* = 0.37, 95% CI: -0.51, 1.25**).

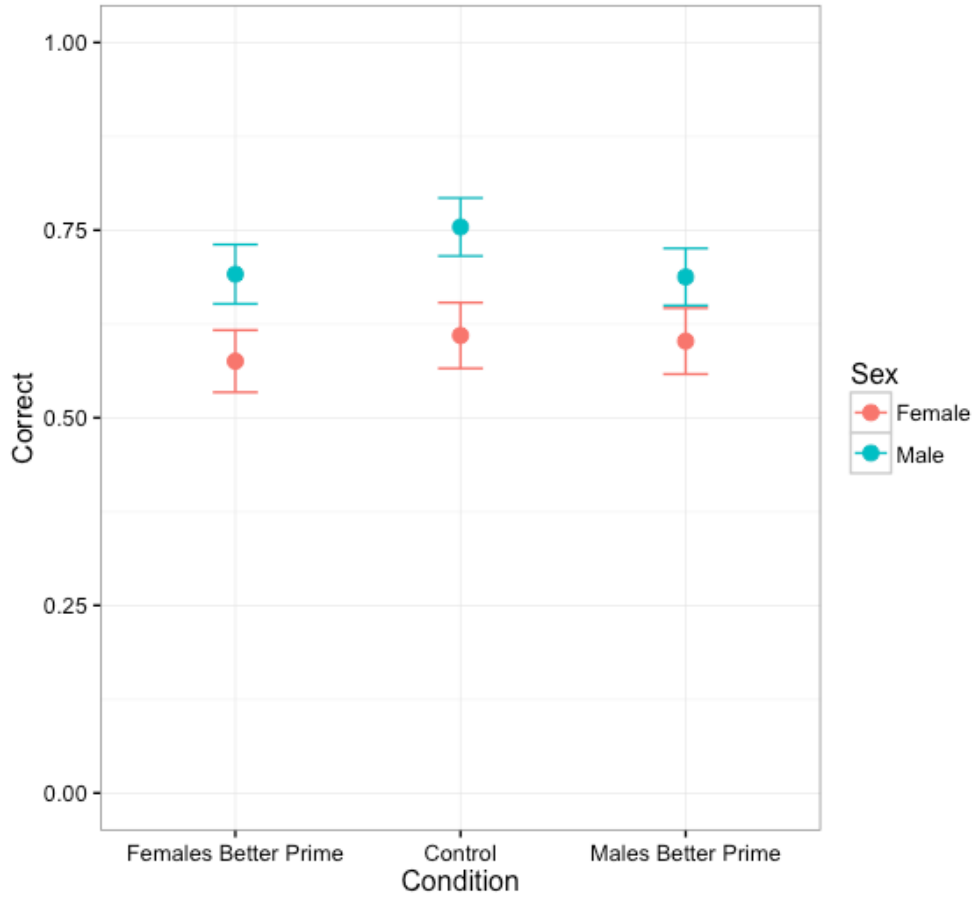


***Figure 3.5**: Plot showing men and women's mean proportion of correctly answered trials per condition. Error bars show 95% confidence intervals.*

3.1.9 Mediation analysis

As stated above, neither condition predicted performance directly, however the 'females better' condition did affect participants' confidence compared to the control, and confidence predicted performance **(β = 0.22, 95% credible interval: 0.15, 0.28)**. Therefore I conducted a mediation analysis to see if there was an indirect effect of the 'females better' condition on performance, mediated by confidence. There was evidence of a small indirect effect of the 'females better' condition on performance, mediated by confidence **(β median = 0.09, 95% credible interval: 0.02, 0.18)**, which suggests that the raised confidence of participants in the 'females better' condition had an indirect effect on their performance in the task, despite no detectable direct effect of condition on performance.

3.1.10 Manipulation check

Our manipulation check data supported our main results; only the 'females better' condition had an effect on how men and women expected to perform compared to the opposite sex, in that both men and women reported that they expected to perform better than the opposite sex in the 'females better' condition compared to the control **(β = 0.32, 95% CI: 0.07, 0.58)**, but not in the 'males better' condition **(β = 0.32, 95% CI: -0.32, 0.89)**. There was no evidence for sex affecting the likelihood of remembering the task manipulation **(β = 0.36, 95% CI: -0.86, 1.59)**, however more participants remembered reading the 'females better' manipulation than the 'males better' manipulation **(β = -1.15, 95% CI: -1.96, -0.45)**. As well as being more memorable, there was evidence that participants were less likely to believe the 'females better' manipulation than the 'males better' manipulation **(β median = -1.00, 95% CI: -1.75, -0.34)**, despite it increasing their confidence.

3.1 Discussion

In this online study, I manipulated participants' confidence in a mental rotation task by using a stereotype threat prime. Both men and women were more confident in the condition in which they were told that women perform better within the instructions of the task. This result supports the idea that confidence plays a role in the stereotype threat effect on mental rotation performance, and supports a previous finding in which women's confidence was lower than men's on a mental rotation task but not on a letter transformation task (Cross et al. 2016). Moreover, confidence, but not performance was affected by the stereotype threat prime, which is in line with the idea that changes in performance found in other stereotype threat studies on mental rotation could be a downstream effect of changes in confidence. In this study, men had higher performance than women in the control condition, but performance was not affected by condition.

Unexpectedly, when told that women perform better on the task, men's average confidence rating was also higher than those in the control condition, and the 'males better' condition. This was unexpected in that I predicted men's confidence to be higher in the 'males better' condition compared to the control, and compared to the 'females better' condition. However, some previous research has also found men's performance to increase in response to a stereotype threat (Hirnstein et al. 2012) and has been termed "stereotype reactance" (Nguyen & Ryan 2008). Although men's increase in confidence seems counterintuitive, studies measuring the effect of stereotype threat on men's performance so far have given mixed results. Men's performance has been reduced in social and emotional intelligence tests when given stereotype threat manipulations (Koenig & Eagly 2005; Leyens et al. 2000), as well as a verbal fluency task (Hirnstein et al. 2014). However, this result was in direct contrast with results from a study using the same task and same manipulation in which men's performance did not change (Hausmann et al. 2009), as well as a study in which men's performance increased in response to the stereotype manipulation (Hirnstein et

al. 2012).

One way of interpreting such mixed results is in terms of how each participant appraises or interprets the manipulation. Hausmann (2014) has suggested that reference to a stereotype could be interpreted as challenging and could therefore result in motivation to compete, resulting in enhanced performance. Alternatively the stereotype could be perceived as threatening in which an arousal threshold is reached meaning that cognitive deficits occur, resulting in decreased performance. This interpretation fits within the Yerkes-Dodson framework of an inverted U-shape relationship between arousal and cognition (Yerkes & Dodson 1908). Our manipulation directly told participants that either men or women were better at the task, and could therefore be interpreted as inducing competition, enhancing motivation and consequently increasing performance in men. More evidence that competition may play a role in differing appraisals of stereotype threat comes from a study in which testosterone has been measured in the context of stereotype threat. The results demonstrated an increase in testosterone levels in men in the stereotype activated condition on mental rotation tasks compared to males in the control condition (Hausmann et al. 2009). Previous studies have also shown that fluctuations in testosterone levels can be induced by competition in men and women (Bateup et al. 2002; Archer 2006). More generally, research indicates a tendency for men to be more competitive than women, and that this difference persists throughout childhood (Sutter & Glätzle-Rützler 2014). In our study, we found no direct effect of condition on performance, but an indirect effect mediated by confidence. This result supports the idea that stereotype threat effects on performance may be a downstream effect of motivational or competitive responses to the primes, which have a direct effect on confidence first and foremost.

Strikingly, in our study there was no difference in confidence between controls and the ‘males better’ condition. Indeed, telling participants that men perform better on the task had no discernible effect on either men or women’s confidence. This response is indicative of

both men and women assuming that males are better at this task, even when no information is provided to encourage this, and no mention of gender is made in the task. This result has some potentially important implications, especially in the context of debates around women in STEM disciplines (e.g. Smith 2011; Blickenstaff 2005). If men and women approach certain tasks with the assumption that men are better, and this is reflected in their confidence levels, this could have important downstream effects for women leading to either suboptimal performance or disengagement with the area altogether (e.g. Bench et al 2015; Rothgerber & Wolsiefer 2014). Encouragingly, simply stating that the task relies on skills in which women outperform men increases women's (and men's) confidence significantly, at least in this task, in this context.

Study 2: Confidence and Conformity

Having successfully manipulated participants' confidence levels in Study 1, I then ran the same study with an additional social learning stage, to see if these confidence changes had a predictable effect on participants' conformity levels.

I ran the same study as Study 1, this time in person rather than online, and with the two opposing conditions only; the 'males better' and 'females better' primes. This time participants were shown social information after each trial and were given the opportunity to change their answer after viewing the social information. Based on the results from the pilot study in Study 1, I predicted that participants in the 'females better' condition would have higher confidence than participants in the 'males better' condition, and that higher confidence in this condition would result in lower levels of conformity compared to the 'males better' condition.

3.2 Methods

3.2.1 Procedure

The procedure, materials and primes were identical to Study 1 (see sections 3.1.2, 3.1.3 and 3.1.4) except for the addition of social information (see below) and except that no Control condition was used. In Study 1, the control condition produced indistinguishable levels of confidence to the 'males better' condition, thus I just ran the two priming conditions in Study 2. As it takes longer to recruit participants in person than it does through an online platform such as Qualtrics, having only two conditions allowed me to recruit a similar number of participants for each condition as the online study.

3.2.2 Participants

101 participants, 50 males and 51 females, (Condition 1: 24F, 26M. Condition 2: 27F, 24M) were recruited using the University of St Andrews SONA system and posters advertising the experiment around the university campus. The experiment was conducted in person in the Psychology department at the University of St Andrews. All participants had to read an information page and give their full consent before taking part in the experiment. The experimental protocol was approved by the departmental ethics committee (see Appendix). 94% of participants fell into the 16-30 range and 6% in the 31-60 range. Participants were allocated to groups when possible, with a minimum group size of 4 and a maximum of 9, in order to give the impression of the social information being valid and trustworthy and to encourage participants to attend to the social information. All participants were reimbursed £3 for taking part in the study and were offered a bonus payment of £2 for achieving a high score in the experiment.

3.2.3 Social information

After participants gave each answer and their confidence rating, identically to Study 1, they were reminded of their response, and shown the responses of other participants (see **Fig. 3.6**) The social information was displayed as 12 circles, representing a majority “yes” or “no” response. The participants were then asked “Now, what do you think? Did the images match?” and were given the opportunity to change their mind or stick with their original answer.

The social information was generated from answers in Study 1, and also generated by me to ensure that the information always portrayed a majority answer (i.e. never 6:6). Participants saw genuine social information from Study 1 for the first three questions, and the rest was generated for the experiment. Participants were motivated to attend to the social information

by being told that only their second answer, after viewing the social information, would count towards their score. Participants were motivated to achieve a high score by being offered a bonus payment of £2 if they get a certain number of questions right. Participants were told during the debrief at the end of the experiment that their payment was only based on the first three questions in which the social information was genuine, to avoid participants feeling as if they had been deceived and that the deception may have altered their payment.

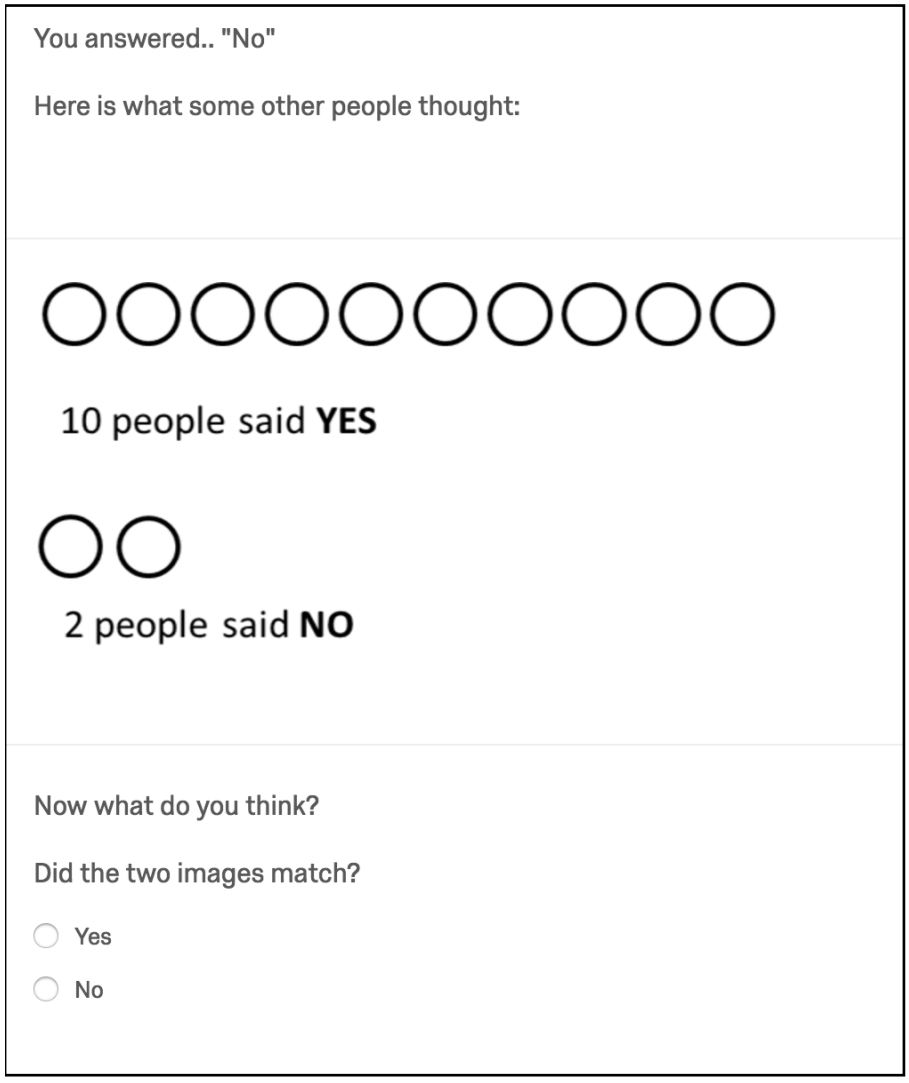


***Figure 3.6**: Example of social information shown to participants on screen after each trial*

3.2.4 Statistical Analysis

The data were analysed in an identical way to Study 1 (see section 3.1.6) except only one parameter is used to compare the effects of each condition to each other, as no control condition was used as the baseline in this study.

To model “switching” i.e. when participants’ second answer differed to their first, the probability that a participant’s answer changed was modelled as a Bernoulli variable (switched = 1, and did not switch = 0) using a logit link function, identical to the previous models. The binomial logistic regression model included a baseline value, an effect for sex, condition, interaction between sex and condition, an effect for confidence, an interaction between sex and confidence, an effect for number of demonstrators disagreeing, and two random effects to allow for variation between individuals and between questions.

A bayesian mediation analysis (Yuann & Mackinnon 2009) was used to to assess the indirect effect of sex on switching via confidence.

All plots are a display of the raw data only, with 95% confidence intervals. All results in brackets are parameter estimates from the model output, including the median and 95% credible intervals. 95% credible intervals that include zero suggest there is little evidence that the parameter had an effect on the outcome variable. 95% credible intervals that do not include zero suggest that the parameter had a negative or positive effect on the outcome variable in the model, depending on the direction of the estimates.

3.2 Results

3.2.5 Confidence

Sex predicted confidence, in that men gave higher confidence ratings than women overall **(*β* = -0.57, 95% CI: -0.90, -0.28, see Fig.3.7)**. In contrast to Study 1, condition did not predict confidence in that confidence levels were similar across the two conditions **(*β* = -0.07, 95% CI: -0.39, 0.25**). There was no interaction between sex and condition ***(β* = -0.11, 95% CI: -0.73, 0.52)**. Performance predicted confidence, in that participants gave higher confidence ratings when they answered a question correctly compared to when they answered a question incorrectly **(*β* = 0.58, 95% CI: 0.48, 0.69, see Fig.3.8)**.

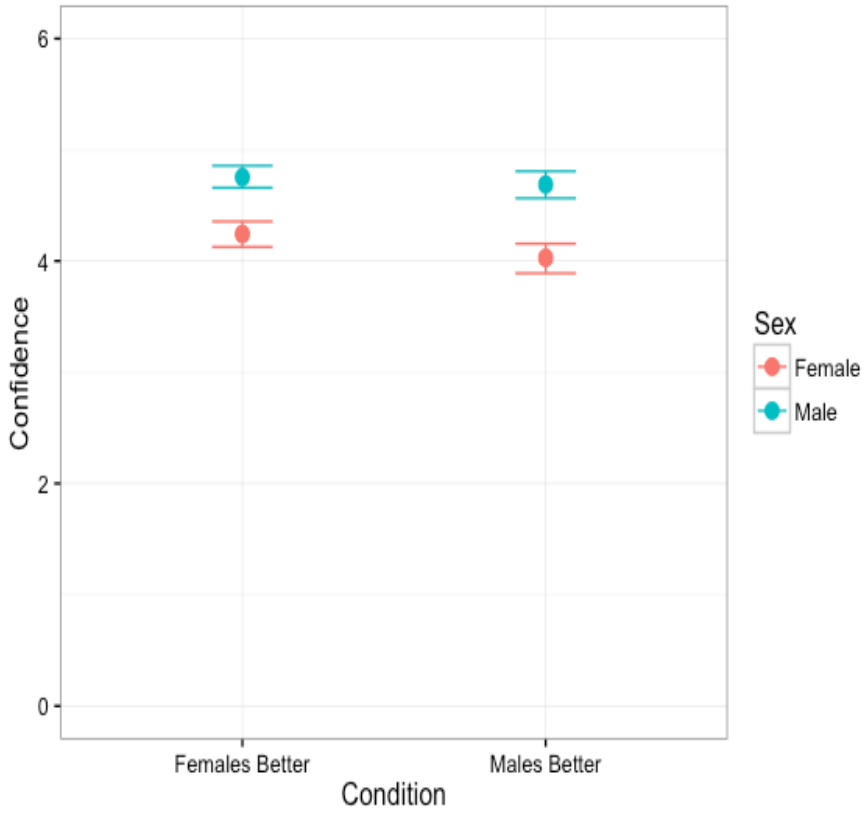


***Figure 3.7**: Plot showing men and women's mean confidence ratings in each condition. Error bars show 95% confidence intervals, bootstrapped for non-normal data.*

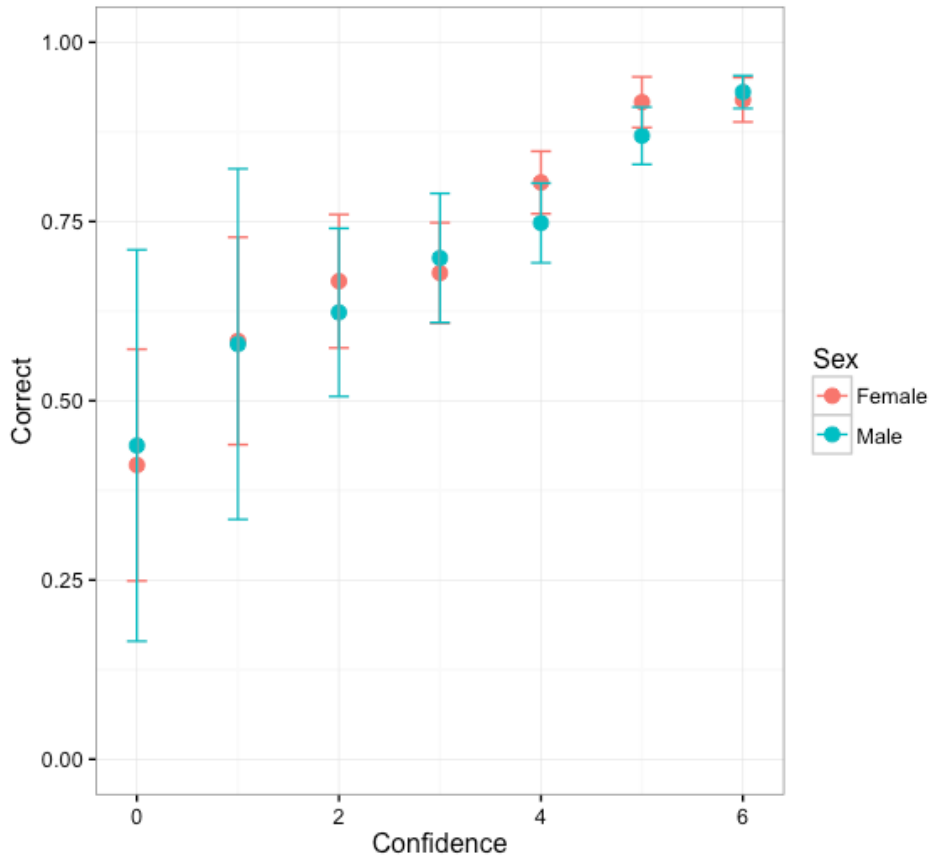


***Figure 3.8**: Plot showing the relationship between men and women's confidence ratings and their proportion of correct trials. Error bars show 95% confidence intervals.*

3.2.6 Performance

Unlike Study 1, sex did not predict performance, in that men and women gave a similar proportion of correct answers in the task **(*β* = -0.17, 95% CI: -0.59, 0.25, see Fig.3.9)**. Also in contrast to Study 1, condition did predict performance, in that participants performed better in the 'females better' condition compared to the 'males better' condition **(*β* = -0.43, 95% CI: -0.86, -0.02, see Fig.3.9)**. There was no evidence of an interaction between sex and condition **(Condition 1: *β* = -0.52, 95% CI: -1.35, 0.30)**.

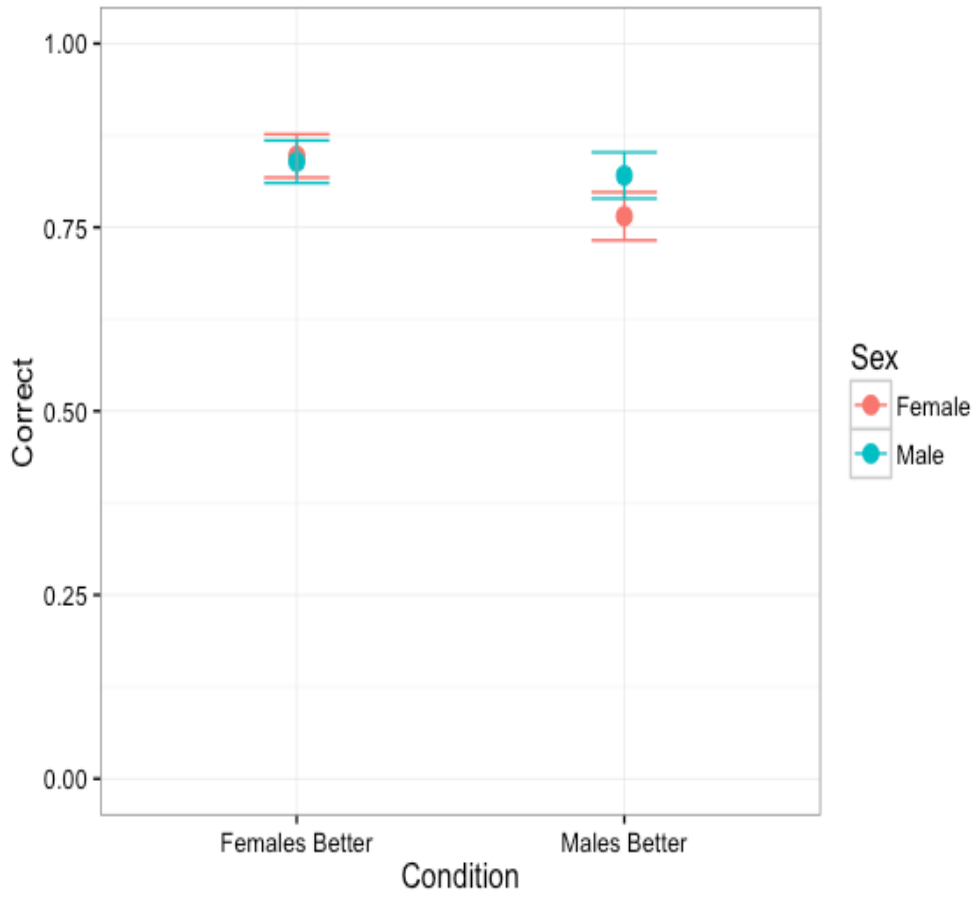


***Figure 3.9**: Plot showing men and women's mean proportion of correctly answered trials for each condition. Error bars show 95% confidence intervals.*

3.2.7 Switching

Because the social information was randomly generated for the majority of questions, and as participants could choose either answer before seeing the social information, participants did not see a majority of demonstrators disagreeing with them for all trials. Therefore a breakdown of social information is provided in **table 3.1** below. Out of 2419 total trials, the majority disagreed with participants on 1138 trials altogether. Out of these trials in which the majority disagreed with participants, the majority was correct (and thus the participant incorrect) on 260 trials. Participants were more likely to switch answer to conform to an opposing majority, as 95% of all switching occurrences were within these trials. Out of all

trials, participants only switched 8% of the time. However, in trials in which the majority disagreed, they switched 17% of the time. In trials in which the majority disagreed and the majority were correct (260 trials) participants switched 29% of the time.

Trial Type	Total Trials	Total Switched	Proportion Switched
All	2419	204	0.08
Majority Disagreed	1138 (m=50, f=51)	194 (m=16, f=13)	0.17
Majority Disagreed AND Majority Correct	260 (m=41, f=43)	75 (m=13, f=11)	0.29
Majority Disagreed AND Majority incorrect	878 (m=50, f=51)	119 (m=13, f=10)	0.16

Table 3.1: *Table showing a breakdown of the social information displayed on all trials across all participants.*

All results given below are from the statistical models described in section 3.2.4 which included all trials. Confidence predicted switching, in that when participants gave lower confidence ratings they were more likely to switch answers after seeing social information **(β = -1.02, 95% CI: -1.20, -0.86, see Fig 3.10)**. Performance did not predict switching **(β = -0.17, 95% CI: -0.63, 0.31)**. Number of demonstrators disagreeing with the participant did predict switching, in that more demonstrators disagreeing led to higher rates of switching answer **(β = 0.47, 95% CI: 0.39, 0.55, see Fig 3.11)**. Sex did not predict switching **(β = 0.37, 95% CI: -0.86, 1.64).** There were no interactions between sex and condition, or sex and confidence **(β = -0.24, 95% CI: -1.80, 1.30; β = -0.08, 95% CI: -0.38, 0.22).**

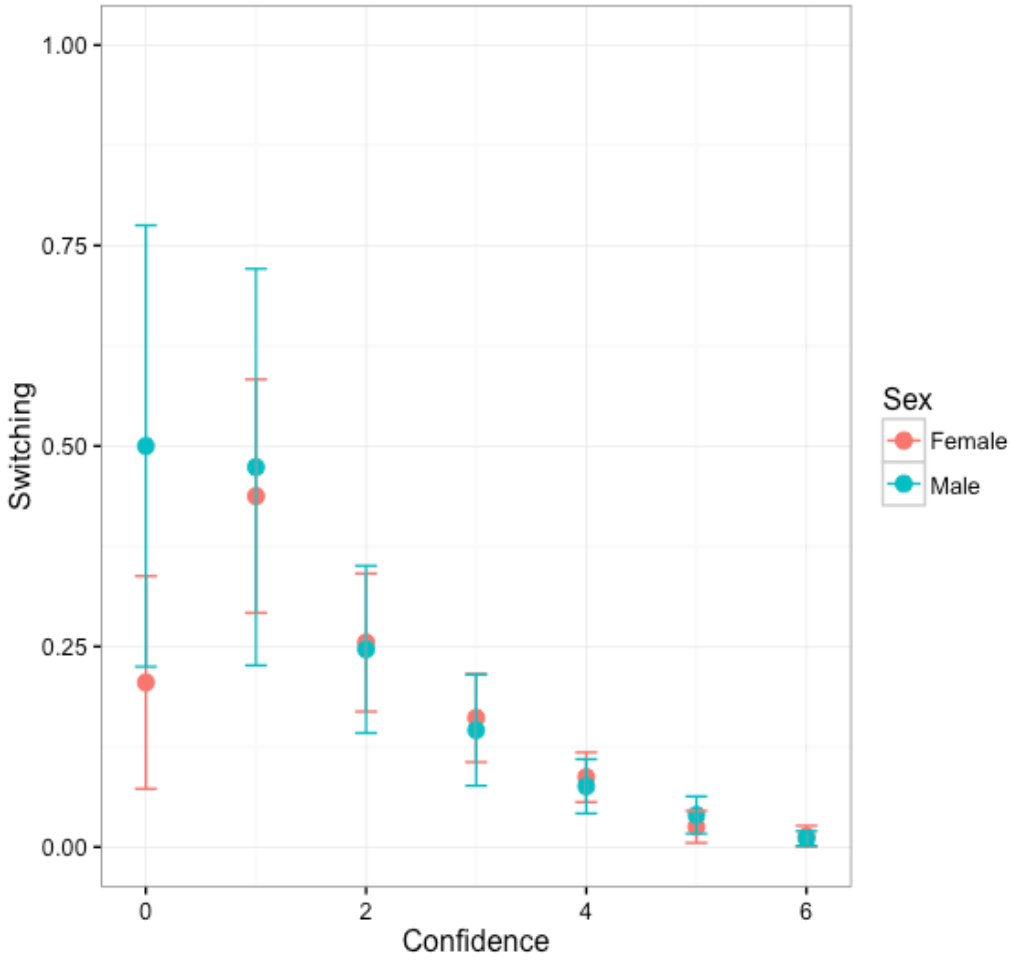


***Figure 3.10**: Plot showing the relationship between men and women's mean confidence ratings and the proportion of trials they switched their answer on. Error bars show 95% confidence intervals.*

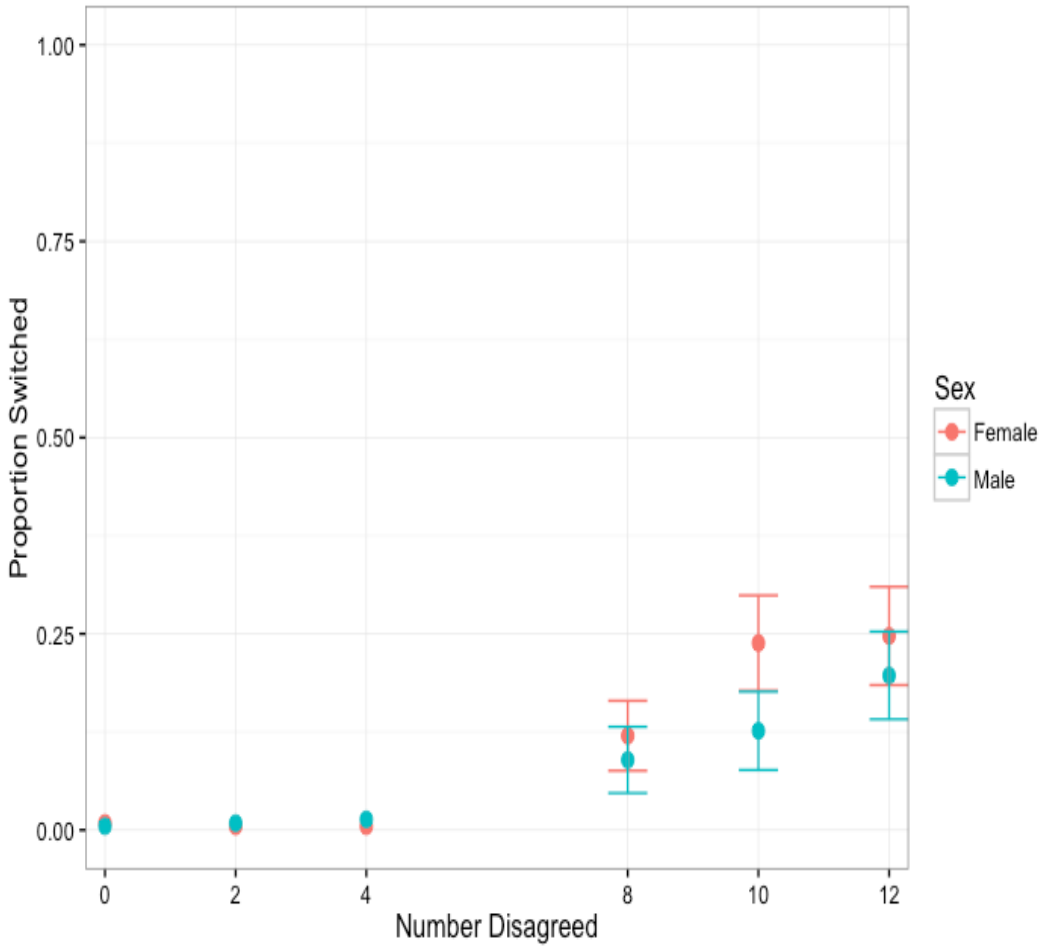


Figure 3.11: *Plot showing the relationship between number of demonstrators disagreeing with the participant and proportion of trials on which they switched their answer after seeing the demonstrators' answers. Error bars show 95% confidence intervals.*

3.2.8 Mediation analysis

As stated above, sex did not directly predict switching, although it did predict confidence, and confidence predicted switching. Therefore I conducted a mediation analysis to see if there was an indirect effect of sex on switching, mediated by confidence. The mediation analysis revealed a medium to large indirect effect of sex on switching via confidence **(β = 0.61, 95% CI: 0.26, 0.97)**, which suggests that lower average confidence in women led to a greater probability of switching, despite the lack of a direct effect of sex on switching.

3.2. Discussion

In Study 2, I found that men and women's confidence was not altered by the stereotype threat primes, but men's average confidence remained higher than women's in both conditions. Men and women performed similarly on the task in terms of proportion of correct trials, but those in the 'females better' condition performed slightly better than those in the 'males better' condition. Men and women did not differ in their levels of switching answer, but similarly to Cross and colleagues (2016), I also found an indirect effect of sex on switching answer, mediated by confidence. My methodology was similar to Cross and colleagues' (2016), except that I used a different version of the mental rotation task in which bodies rather than blocks were rotated. Sex predicted confidence in this version of the task also, in that men were more confident than women. Similarly, confidence predicted likelihood of switching answer, as did the number of demonstrators disagreeing with the participant, as was the case in Cross and colleagues' study.

The results of this study therefore support previous findings suggesting that conformist behaviour may sometimes be underpinned by a copy-when-uncertain social learning strategy (Morgan et al. 2012; Cross et al. 2016), in that participants that were less certain in their answer were subsequently more likely to switch their answer after seeing a majority of demonstrators disagree with them. Similarly to previous studies, participants were more likely to switch their answer as the number of demonstrators disagreeing with them increased. On this version of a mental rotation task in which bodies rather than cubes were used, men still had higher confidence than women on average. This resulted in an indirect effect of sex on switching, mediated by confidence, even though no direct effect of sex on switching was found.

Although there was a sex difference in confidence on this task, there wasn't a corresponding sex difference in performance. Also In contrast to Study 1, confidence was not affected by

the stereotype threat primes in either condition. One reason for these conflicting findings may be the differences in sample between Study 1 and Study 2. Study 1 was conducted online, and so the sample of participants may have been more varied and diverse compared to the participants from Study 2, who were recruited predominantly from the University. University students could be expected to have higher confidence levels in the task in general, due to regular tests and performance assessments throughout university. In contrast, an average member of the public may be less familiar with this type of task, as well as having to have their cognitive performance assessed in this exam-style manner. It is also possible that they were more affected by the information they were told about the task beforehand if they were less familiar with the task to begin with. University students may be less likely to believe primes in an experiment due to a greater exposure to the idea of priming in experimental settings. A way to test this would be to re-run Study 2 with an online sample to see if the confidence difference found in Study 1 is replicated in another online sample (see Study 3).

Another possibility is the presence of an experimenter effect in Study 2. In a previous study (Morgan et al. 2012, personal communication) the presence of a male experimenter compared to a female experimenter led to lower female confidence ratings. As I was the only experimenter present in each lab study, this could have affected female confidence levels to be higher than it would be if it were a male experimenter. Again, this could be explored by re-running Study 1 with another online sample in which an experimenter not be present or visible to the participants.

A particularly counter-intuitive finding from this study is that, in contrast to Study 1, confidence was not affected by condition, but performance was. Participants performed slightly better in the ‘females better’ condition compared to the ‘males better’ condition. In Study 1, participants in the ‘females better’ condition had higher confidence than those in the ‘males better’ condition, but not higher performance. The result of Study 1 supported my

prediction that stereotype threat affects confidence first and foremost, and any performance effects are a downstream consequence of confidence differences. However, in Study 2, participants' performance appears to be altered by the stereotype threat condition, but not their confidence. One potential explanation could be related to the fact that the student participants are performing better overall, with their mean correct trials at 82% compared to 65% correct in the online sample from Study 1. Furthermore, when looking at the raw data, the student sample appears to show a stronger correlation between performance and confidence than the online sample does (see **Figs 3.4 and 3.8**). Thus, the student sample appears to have a more accurate assessment of their performance on the task than the online sample. It is therefore possible that as the student sample was both confident and capable at the task, the 'females better' condition enabled the students to perform slightly better without affecting their ability to rate their confidence accurately. In Study 1, however, women were slightly less accurate at rating their confidence in that women who gave the highest confidence ratings (5 and 6) were not performing any better than women who gave a confidence rating of 4. Thus, perhaps the 'females better' condition falsely boosted female confidence in Study 1 whilst not improving their performance. However, these explanations are highly speculative and more data would be needed to support them.

Study 3: Replication of Study 2

As the results of Study 1 were not replicated in Study 2, I wanted to repeat Study 2 with an online sample, to address the fact that the stereotype threat manipulations may have failed due to the sample in Study 2 comprising of predominantly undergraduate students.

3.3 Method

3.3.1 Procedure

The procedure is identical to Study 1, this time with the social information of Study 2 at the end also (see sections 3.1.2, 3.1.3 and 3.2.3).

3.3.2 Participants

130 participants were recruited altogether, 65 male and 65 female. (Control: 22M, 22F. Condition1: 22M, 22F. Condition 2: 21M, 21F). Participants were recruited using Qualtrics, as in Study 1. Participants knew that they could withdraw from the study at any time without explanation. The study was approved by the departmental ethics committee for the School of Psychology & Neuroscience at the University of St Andrews. The sample requested was from the UK, aged over 16. Participants were randomly assigned to the control and two conditions. 38% of participants fell into the 31-60 years age category, 32% in the 16-30 range, 30% in the 61-80 category and 1 participant preferred not to disclose their age category. Participants were reimbursed £3 for their time and given the opportunity to earn a £2 bonus based on performance.

All plots show the raw data only, with 95% confidence intervals. All results in brackets represent model parameter estimates, including 95% credible intervals. 95% credible intervals that include zero suggest there is little evidence that the parameter had an effect on the outcome variable. 95% credible intervals that do not include zero suggest that the

parameter had a negative or positive effect on the outcome variable in the model, depending on the direction of the estimates.

3.3 Results

3.3.3: Confidence

In contrast to Study 1, but similarly to Study 2, there was no strong evidence that either condition affected confidence ratings compared to the control condition **(Condition 1: β = 0.21, 95% CI: -0.22, 0.63, Condition 2: β = 0.07, 95% CI: -0.36, 0.50 see Fig 3.12)**. As was the case in both Study 1 and Study 2, sex predicted confidence: women gave lower confidence ratings than men on average across all three conditions **(β = -0.42, 95% CI: -0.76, -0.07).** Consistently with the previous two studies, there was no interaction between sex and either condition **(Condition 1: β = 0.29, 95% CI: -0.57, 1.16, Condition 2: β = -0.06, 95% CI: -0.93, 0.81)**. Consistently with the previous two studies, performance predicted confidence, in that participants gave higher confidence ratings when they answered a question correctly compared to when they answered a question incorrectly ***(β* = 0.15, 95% CI: 0.07, 0.23, see Fig 3.13).**

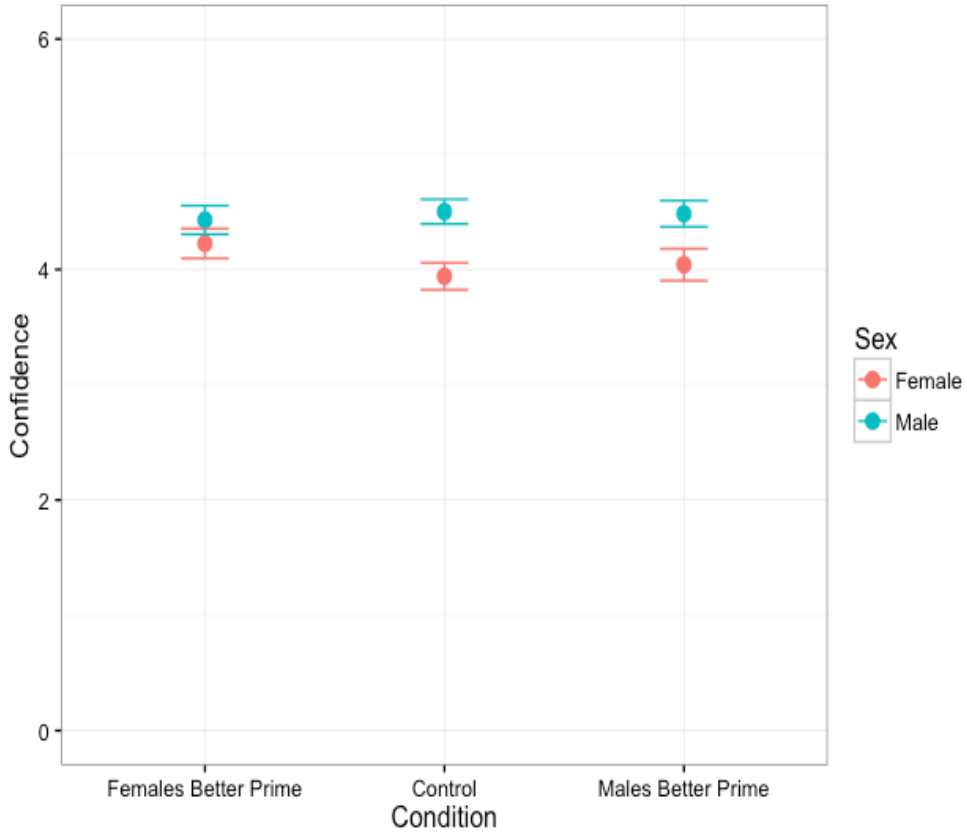


***Figure 3.12**: Plot showing mean confidence ratings per sex, per condition. Error bars are 95% confidence intervals bootstrapped for non-normal data.*

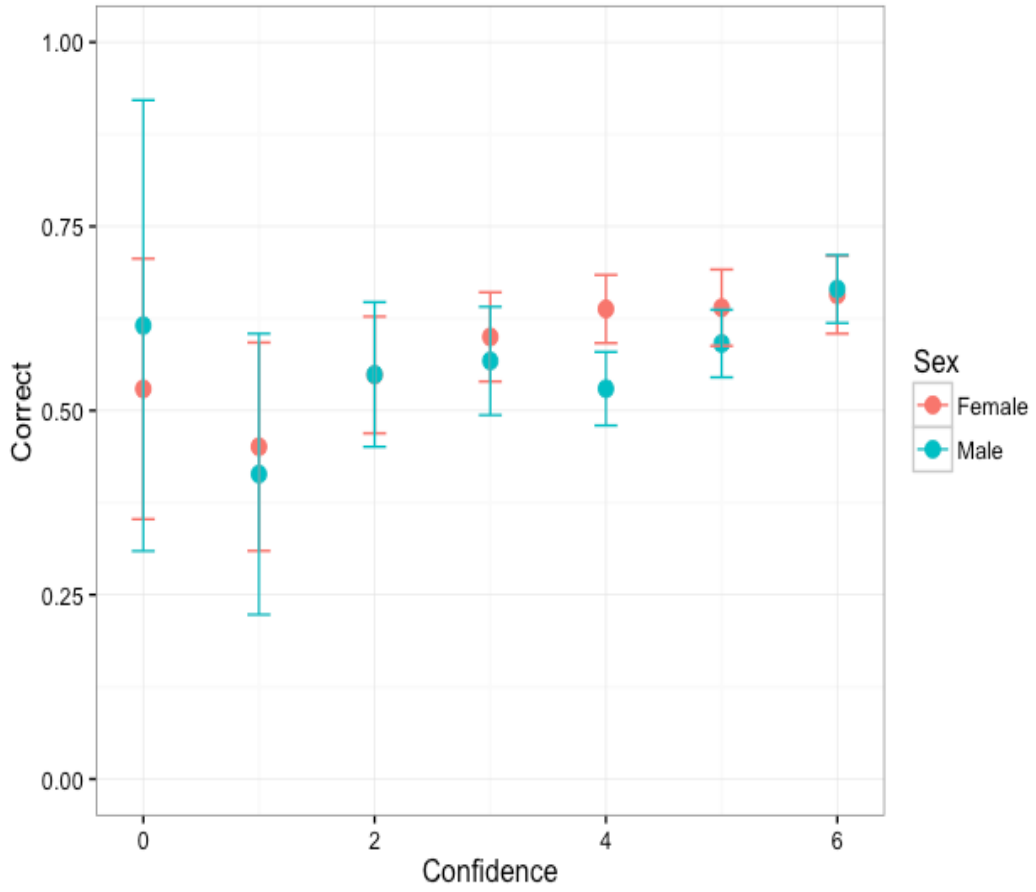


***Figure 3.13**: Plot showing the relationship between confidence ratings and proportion of correct trials per sex.*

3.3.4 Performance

In contrast to Study 1, but similarly to Study 2, sex did not predict performance in the control **(β = 0.41, 95% CI: -0.30, 0.93, see Fig.3.14),** or across all three conditions overall **(β = 0.16, 95% CI: -0.16, 0.48).** In contrast to Study 2, but similarly to Study 1, there was no strong evidence that either condition affected performance compared to the control **(Condition 1: β = 0.20, 95% CI: -0.18, 0.52, Condition 2: β = 0.10, 95% CI: -0.38, 0.44 see Fig.3.14).** As was the case in both previous studies, there was no evidence of an interaction between sex and condition **(Condition 1: β = -0.19, 95% CI: -0.99, 0.48, Condition 2: β = -0.37, 95% CI: -1.11, 0.20).**

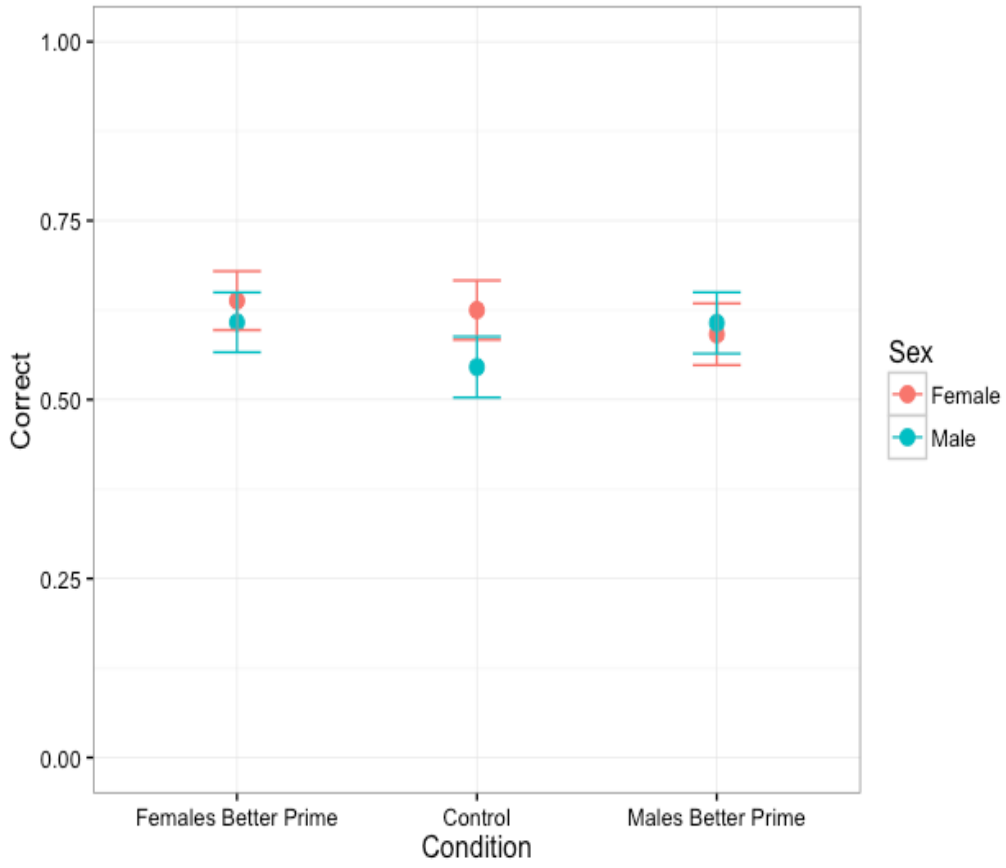


***Figure 3.14**: Plot showing the mean proportion of correct trials per condition per sex, error bars show 95% confidence intervals.*

3.3.5 Switching

As the social information was randomly generated for the majority of questions, and as participants could choose either answer before seeing the social information, participants did not see a majority of demonstrators disagreeing with them for all trials. Therefore a breakdown of social information is provided in **table 3.2** below. Out of 3120 total trials, the majority disagreed with participants on 1524 trials altogether. Out of these trials in which the majority disagreed with participants, the majority was correct (and thus the participant incorrect) on 706 trials. Participants were more likely to switch answer to conform to an

opposing majority, as 93% of all switching occurrences were within these trials. Out of all trials, participants only switched 15% of the time. However, in trials in which the majority disagreed, they switched 28% of the time. In trials in which the majority disagreed and the majority were correct (706 trials) participants switched 34% of the time.

Trial Type	Total Trials	Total Switched	Proportion Switched
All	3120	462	0.15
Majority Disagreed	1524 (m=65, f=65)	428 (m= 24, f=28)	0.28
Majority Disagreed AND Correct	706 (m=65, f=64)	241 (m=24, f=27)	0.34
Majority Disagreed AND Incorrect	818 (m=65, f=65)	187 (m=16, f=14)	0.23

Table 3.2*: Table showing a breakdown of the social information displayed on all trials across all participants.*

All results shown below are from the statistical models described in section 3.2.4 which included all trials. As was the case in Study 2, confidence predicted switching in that lower confidence ratings increased the likelihood of switching answer **(*β* = -0.93, 95% CI: -1.05, -0.70, see Fig.3.15).** In contrast to Study 2, performance also predicted switching, in that lower performance on the task increased the likelihood of switching answer **(*β* = -0.48, 95% CI: -0.77, -0.13).** As the number of demonstrators disagreeing with the participant increased, the likelihood of them switching answer increased **(*β* = 0.34, 95% CI: 0.30, 0.39, see Fig. 3.16).** There was no sex difference in the likelihood of switching overall **(*β* = -0.52, 95% CI: -1.62, 0.66).**

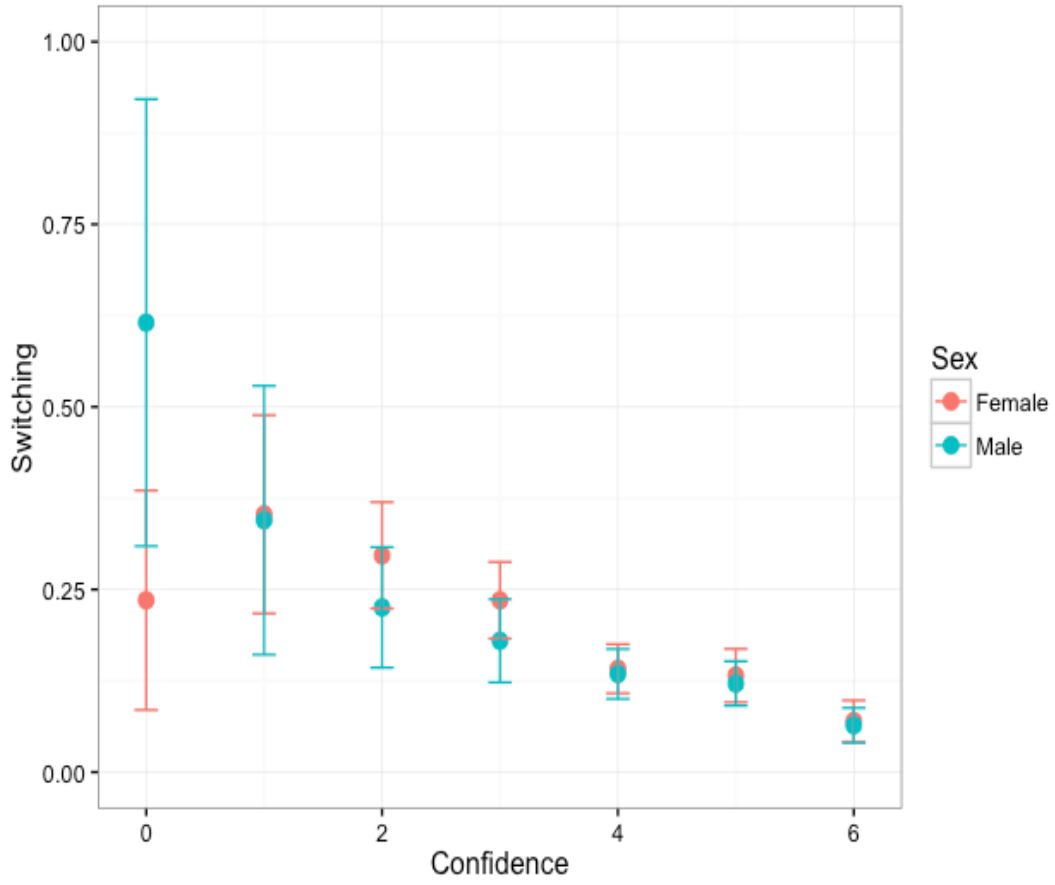


***Figure 3.15**: Plot showing the relationship between confidence and the proportion that switched their answer, per sex. Error bars show 95% confidence intervals.*

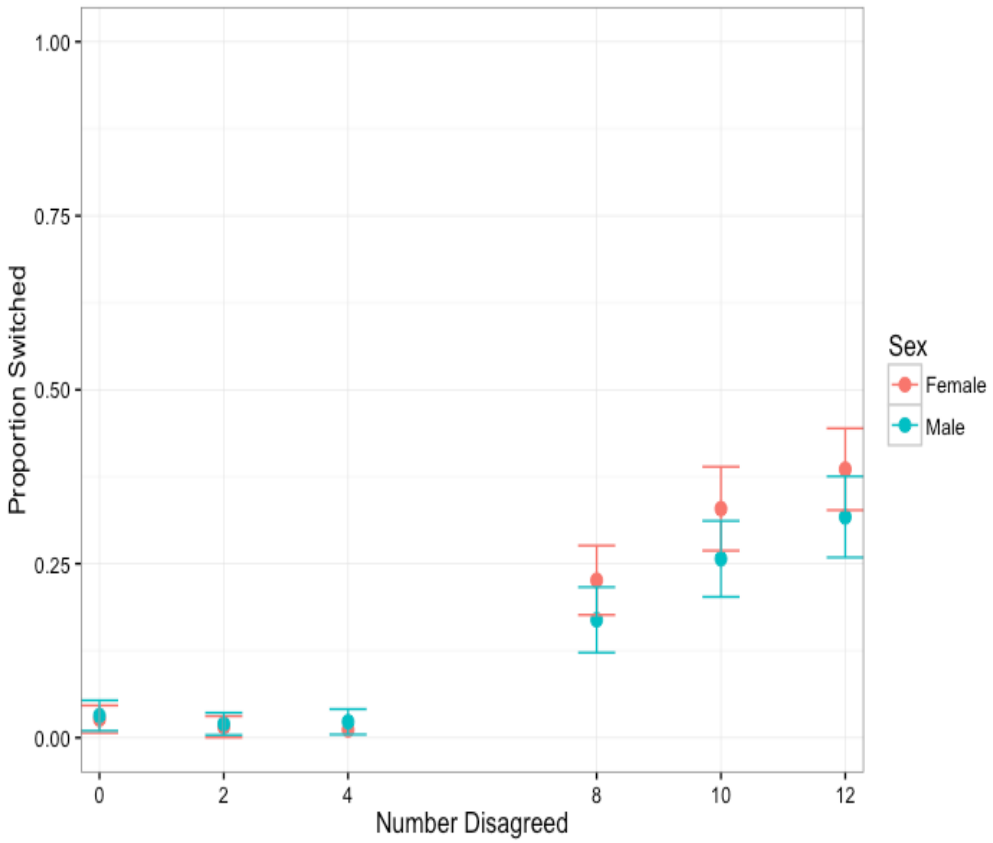


***Figure 3.16**: Plot showing the relationship between the number of demonstrators disagreeing with the participant, and the likelihood of the participant switching answer.*

3.3.6 Mediation analysis

As was the case in Study 2, sex did not predict switching directly, although it did predict confidence, and confidence predicted switching. Therefore I conducted a mediation analysis and found an indirect effect of sex on switching, mediated by confidence. **(β = 0.26, 95% CI: 0.04, 0.49).**

3.3 Discussion

In Study 3, I found that neither 'females better' nor 'males better' conditions affected participants' confidence or performance on the task. Men were overall more confident than women, but there was no sex difference in performance on the task. Consistently with Study 2, participants were more likely to switch their answer when they had lower confidence compared to when they had higher confidence, as well as being more likely to switch answer as the number of demonstrators disagreeing with them increased. As was the case in Study 2, sex predicted confidence in that men were more confident than women overall, and there was an indirect effect of sex on switching, mediated by confidence.

The switching results from Study 3 were consistent with those from Study 2, and supported previous findings (Cross et al. 2016; Morgan et al. 2012), in that lower confidence led to higher levels of switching. This result appears to be robust, at least in the context of mental rotation tasks, and further supports the idea of a copy-when-uncertain social learning strategy in adult humans. Furthermore, as the number of demonstrators disagreeing with the participant increased, so did the likelihood of switching, supporting theoretical and experimental evidence of adaptive conformist transmission in humans (Boyd & Richerson 1985; King & Cowlishaw 2007; Morgan et al. 2012). The fact that women appear to consistently have lower confidence in this task, as well as the classic version of the mental rotation task, and that this low confidence has a mediating effect on social information use, contributes to the social learning literature and merits further study. It seems likely that in cases in which women have lower confidence than men, this confidence difference will have an indirect effect on their social information use.

However, the results of Study 3 did not provide evidence of any effect of the stereotype threat primes, as neither condition affected participants' confidence or performance. It is unlikely that the lack of effect of condition is due to differences in sample, as Study 3's sample was collected from the same resource as Study 1, in that a UK based sample of

adults was requested from Qualtrics. Indeed the spread of age categories of Study 1 and Study 3 were more similar to each other than Study 2. Moreover, the stereotype threat primes and conditions were identical to Study 1, yet neither condition affected participants' confidence. Nor was performance affected by either condition. Thus it is possible that the previous effects of stereotype threat from Study 1 and Study 2 were coincidental and not reflective of an underlying stereotype threat effect induced by our experimental conditions. Another possibility is that our primes were either too subtle or unclear to have a reliable effect, or that the stereotype threat effect is moderated by other factors that I have not considered in these three studies. Possible explanations for the contradictory findings are discussed further below.

General Discussion

Neither of the effects of stereotype threat from Study 1 and Study 2 were replicated in Study 3. In Study 1, participants in the 'females better' condition had higher confidence than participants in the Control or 'males better' condition, but this effect was not present for participants' performance. In Study 2, participants' performance was higher in the 'females better' condition compared to the 'males better' condition, but this effect was not present for participants' confidence. However in Study 3, neither confidence nor performance was affected by the 'females better' or 'males better' condition. Thus, in this replication attempt, I did not replicate either of the effects that were found in the previous two studies.

The inconsistent results from these three studies add to the already contradictory and somewhat confusing nature of the stereotype threat literature. For example, as previously mentioned, stereotype threat has been found to decrease men's performance (Leyens et al. 2000; Koenig & Eagly 2005; Hirnstein et al. 2014), have no effect on men's performance (Hausmann et al. 2009), and increase men's performance (Hirnstein, Freund & Hausmann

2012), known as “stereotype reactance” (Nguyen & Ryan 2008). Various explanations for these mixed results have been put forward relating to how participants are interpreting the stereotype threat primes. One explanation is that some participants interpret the prime as threatening or challenging, which could either increase or decrease motivation (Hausmann 2014). Alternatively, participants could interpret the primes as inducing a competitive context (Hausmann et al. 2009), causing men to respond differently to the primes than women (Sutter & Glätzle-Rützler 2014).

The effects of stereotype threat on women are similarly mixed, with many meta-analyses examining the effect on women’s spatial and mathematical skills in particular (Nguyen & Ryan 2008; Stoet & Geary 2012; Picho et al. 2013; Flore & Wicherts 2014; Doyle & Voyer 2016) as well as one meta-analysis looking at the counter-intuitive “stereotype lift” results in particular (Walton & Cohen 2003). Many moderating effects have been reported, such as task difficulty and prime subtlety (Nguyen & Ryan 2008), level of education and testing context (Picho et al. 2013), sex of experimenter (Doyle & Voyer 2016), plus a confounding effect of adjusted mathematical scores (Stoet & Geary 2012). Furthermore, after controlling for this confounding effect of adjusted scores, Stoet and Geary (2012) only found that 33% of studies replicated the original stereotype threat findings of Spencer, Steele & Quinn (1999). The authors concluded that the evidence did not strongly support stereotype threat as a potential explanation for the gender gap in mathematics.

One potential problem with the stereotype threat literature which is often highlighted is the “file-drawer effect” or publication bias. This is the idea that researchers are more likely to try to publish, or be successful in publishing, studies which find an effect rather than null findings (Rosenthal 1979). In their meta-analysis of stereotype threat effects on girls in maths and spatial skills, Flore and Wicherts (2014) examined the effect of publication bias using several methods. These methods included funnel plot testing (Duval & Tweedie 2000), Egger’s test (Egger, Smith, Schneider & Minder 1997), rank correlation tests (Begg and

Mazumdar 1994) and Ioannidis and Trikalinos' test (Ioannidis & Trikalinos 2007), among others. The authors concluded that there was strong evidence of publication bias in the literature (Flore & Wicherts 2014).

A recent meta-analysis by Doyle and Voyer (2016) examined all four directions of stereotype effects on math and spatial performance; i.e threat to females, threat to males, lift for females and lift for males. The authors examined unpublished data too, but could only access three unpublished datasets, thus were unable to compute a simple comparison of published versus unpublished effect sizes. However, using Egger's test (1997), the meta-analysis revealed a publication bias effect for the threat to females category, but not the other three categories (Doyle & Voyer 2016). Moreover, out of 224 effect sizes, the threat to females category was the only category in which a mean effect size greater than zero was found (Doyle & Voyer 2016).

In Studies 1, 2 and 3 of this chapter, the stereotype threat primes can be interpreted as fulfilling all four possible stereotype effects on gender, in that the 'females better' prime could be interpreted as a threat to males or a lift for females, and vice versa for the 'males better' prime. Therefore for each study, eight different possible results could have supported the presence of a stereotype effect; four for confidence and four for performance. Thus out of all three studies, there could have been any of 24 results interpreted as evidence of a stereotype threat effect. In reality, 2 results were found in support of an effect, thus a proportion of 0.08 studies. In light of this, I am inclined to conclude that these results were spurious and that there is not strong evidence for stereotype threat effects within my three studies. Given the strong evidence of publication bias in two recent meta-analyses of stereotype threat (Flore & Wicherts 2014; Doyle & Voyer 2016), there is reason to suspect that the stereotype threat effect is a product of publication bias. Furthermore, the results that were found in Study 1 and Study 2 were not in the direction of predictions, and so any explanation for these results relies on "HARKing" (Kerr 1998), in which hypotheses are

derived after the results are known. The costs of harking to the progress of science are thought to outweigh the potential benefits and is therefore widely discouraged (Kerr 1998, Forstmeier, Wagenmakers & Parker 2016).

Chapter 4:

Statistical Comparisons of the Stereotype Threat Literature

4.1 Introduction

Given the inconsistencies and uncertainty in the stereotype threat literature as discussed at the end of Chapter 3, I was interested in exploring methods that could prevent the accumulation of a large, ambiguous literature. By simulating datasets based on the stereotype threat literature, I examine the effect of using different methods of analysis on the development of a field of research. In particular, I was interested in whether this certainty is helped or hindered by different statistical practices that are employed by a field of research. I compared four different analysis methods; a frequentist analysis of variance (ANOVA), a frequentist generalised linear mixed model (GLMM), a Bayesian GLMM using markov chain Monte Carlo estimation (MCMC), (henceforth “BGLMM") and a Bayesian GLMM using MCMC methods and posterior passing (henceforth “PP”).

These four approaches have been chosen as they are representative of current approaches used in the stereotype threat literature, but also include more recently developed methods that are yet to be fully adopted by psychologists. For example, ANOVAs have been widely used in psychology for decades and still represent one of the most commonly used analytic approaches in the field of stereotype threat despite suggestions of their inadequacy for many types of experimental design (Jaeger 2008). A move towards using generalised linear mixed models for categorical and binomial data has been suggested as more appropriate than methods often used by psychologists and ecologists (Jaeger 2008; Bolker et al. 2009). Bayesian approaches have also been suggested as an alternative to null hypothesis significance testing (Howard, Maxwell & Fleming 2000; Bolker et al. 2009; van de Schoot et al. 2014), as null hypothesis testing has been debated and widely criticised (Lee & Wagenmakers 2005; Kruschke 2011; van de Schoot 2012).

One of the fundamental distinctions between Bayesian and frequentist approaches is the use of priors, in which bayesian models can include information about the researcher’s prior beliefs and expectations. These beliefs are then updated based on the data put into the

model. Bayesian approaches are not commonly used in the field of psychology at the moment, and when they are implemented, minimally informative priors are often chosen to avoid unintentionally influencing the results in someway. However, proponents of Bayesian analysis suggest that weakly informative priors are preferable over flat, or uninformative, priors (Gelman 2006; McElreath 2016). I was particularly interested in the use of informative priors in the context of many datasets examining the same effect, and whether the use of priors based on previous datasets can help to discern the presence of a true effect in the population. One way of implementing informative priors is a technique known as "posterior passing" (Beppu & Griffiths 2009) in which the posterior distribution from a previous analysis is used to update beliefs and forms the prior distribution of the next analysis. In particular, the mean and variance of the posterior distribution from one analysis becomes the mean and variance of the prior distribution for the subsequent analysis. I was interested to see whether implementing posterior passing in this way would be beneficial for a field such as stereotype threat, in which many experiments are conducted to examine the same effect.

Thus, I was interested in how a field of literature might develop differently over time if posterior passing is implemented, compared to the current norm of null hypothesis significance testing. To examine the accuracy of our implementation of posterior passing, I also conducted a "mega BGLMM" in which all data from each of the 60 datasets were combined into a single dataset, and a single analysis was performed. In reality, multiple datasets representing multiple experiments are rarely similar enough to be combined and used for one single analysis. Therefore, our "mega BGLMM" represents a "best case scenario" comparison to the performance of posterior passing.

4.2 Methods

4.2.1 Summary

The simulation involved the following four steps which will be explained in detail below: 1) a population of one million individuals was simulated, 2) 60 datasets were simulated; each data set comprised of data taken from 80 participants of the population, 3) summary data from the 60 datasets were saved and analysed using the four different analysis methods, and 4) these three steps were repeated 20 times, and results from the four analyses over the 20 repeats were compared and plotted.

Steps one to four were repeated with five different effect sizes and variance levels of stereotype threat in the population. This resulted in 25 example stereotype threat scenarios, ranging from no effect in the population with no variance, to a medium effect in the population with a large amount of variance. Thus, 25 lots of 20 simulations resulted in 500 simulations altogether, with a total sample size of 2.4 million.

4.2.2 Creating the Population

A population of 1,000,000 individuals was created. Each individual was defined by two values; sex (0 or 1, male or female) and a baseline level of performance at the experimental task, which varied between individuals.

To ensure that the population mean of the baseline could be equal to 0, and that the variance among individuals of each sex was the same for both sexes, we employed the following method: first a random value (x) was drawn from a normal distribution with mean 0 and variance (v), then four individuals were created with the following sex and baseline values: [0, x]; [0, -x]; [1, x]; [1, -x]. This procedure essentially created sets of four individuals, in which each individual had an individual with the opposite baseline rate to them, as well as an opposite member of sex with the opposite baseline to them, to ensure

that the variance between and within sex remains the same within each population. This process was repeated 250,000 times to generate a population of 1,000,000 individuals.

4.2.3 Executing a simulated experiment

Datasets were generated by randomly selecting a sample of 80 individuals (40 women and 40 men) from the population. These 80 individuals were split into a control condition group and a stereotype threat condition group (20 men and 20 women in each group). An average response to 25 trials of a binary choice task is generated for each participant by sampling from a binomial distribution and multiplying the result by 25 to get an overall score for the 25 trials. The likelihood of success in the binomial distribution for the each participant was calculated as follows:

$$p_i = logistic(baseline_i + stereotype\_threat\_effect_i * condition * sex)$$

where “stereotype_threat_effect” determines how much the behaviour of participants of sex 1 change in condition 1, as these reflect females in the stereotype threat condition. The values of the stereotype threat parameter are given as the parameter ‘b_sex_cond’, which reflects the interaction between condition and sex. Only effect sizes of this parameter were varied, to reflect different stereotype threat effect sizes in a population, as discussed below. Note that participants of sex 0 are insensitive to condition, as these reflect males. Likewise condition 0 has no effect, as it represents the control condition.

4.2.4 Data analysis

We perform four methods of analysis on each simulated dataset. First we employ the predominant method of analysis used in the stereotype threat literature, analysis of variance (2 x 2 ANOVA). Average success on the task was subjected to a 2(Sex) x 2(Condition) ANOVA, that included a main effect of sex, a main effect of condition, and an interaction between sex and condition at a significance value of $p<0.05$.

The second method is a generalised linear mixed model (GLMM) with a binomial distribution for the response variable and a logit link function. The same outcome and predictor variables are used as in the ANOVA, but a random effect for participant is implemented. This method fits parameters based on a maximum likelihood approach and estimates the linear effect that our manipulation and independent variables have on the log odds of success in any given trial.

The third approach uses the same GLMM model formulation but uses Bayesian MCMC estimation to produce posterior distributions of the parameter estimates in Jags. Minimally informative priors were implemented across all parameters in the Bayesian GLMM.

The fourth approach is a bayesian GLMM with “posterior passing,” (Beppu & Griffiths 2009) in which the prior for the interaction parameter (i.e. the stereotype threat effect) in each analysis is specified using values taken from the posterior distribution of the interaction parameter estimate from the previous data set in the sequence. As it isn’t possible for specifications of the entire posterior to be implemented within the subsequent analyses, the mean and variance of the posterior distribution are used to define a normal distribution used as the prior in the next analysis. This method is not commonly used to date but has been chosen to demonstrate a way in which posterior passing could be easily implemented in the literature. To test whether summarizing the posterior in this way affected the accumulation of knowledge across multiple experiments, a “mega bGLMM” was performed, in which all datasets were combined and analysed with a single bayesian GLMM. The purpose of this

comparison was to compare our implementation of posterior passing to the greatest possible performance that posterior passing could theoretically achieve. Thus this mega bGLMM can be viewed as an unattainable meta-analysis, in which all datasets and experimental methods are similar enough to be integrated into one single analysis. This scenario is unlikely to ever be the case in reality, unless multiple replication experiments are performed. Hence, rather than propose a 'mega bGLMM' as a form of analysis to be adopted, we are merely using this hypothetical scenario as a means of evaluating the success of posterior passing.

4.2.5 Comparing the Methods

To compare the four methods, parameter estimates and confidence intervals (or credible intervals where appropriate) for all methods (except posterior passing) were calculated and averaged over the 60 datasets. For posterior passing, only the final dataset in the sequence was used. This was to reflect the fact that in posterior passing only, information from each dataset is incorporated into each subsequent analysis, and thus the final analysis produces the estimate of interest. Thus, the averaged estimates of ANOVA, GLMM and bGLMM's represent a summary of their 60 separate attempts at finding the true effect in the population, in a similar vein to a meta analysis. It is these averaged estimates of ANOVA, GLMM and bGLMM that are then compared to a single PP estimate. Thus, although posterior passing also analyses 60 different datasets, we only use its final analysis as a comparison with the other methods.

These results were then used to calculate four summary results; 1) Estimate of effect size, 2) Positive effect rate, 3) Uncertainty around the effect and 4) Error of the estimate

1) The estimates of effect sizes are the model parameter estimates for the interaction parameter only, i.e. the stereotype threat effect. These estimates are averaged across all 20 simulations for the 60 data sets and include 95% confidence intervals.

2) Positive rates refer to how often the analysis finds an effect, i.e. if the confidence intervals exclude zero, then this is evidence of a positive effect.

3) Uncertainty is calculated by subtracting the upper estimate from the lower estimate, to get the relative width of the confidence interval around the effect.

4) Error is calculated by subtracting the estimated effect from the true effect implemented in the population, to get the relative distance of the estimate from the true effect.

4.2.6 Exploring the Parameter Space

We varied the parameter values of the stereotype threat effect (b_sex_cond) to be between 0 and 0.38, to cover the effect sizes reported in meta-analyses of the stereotype threat literature (e.g. Doyle & Voyer 2016).

We varied the baseline variance to be between 0 and 2, to cover a wide range of variation from no variation to large amounts of variation around the baseline response in the population.

All other parameter values were kept constant at zero for all simulations.

4.3 Results

4.3.1 Estimates for each analysis method

The following plots (**Fig.4.1**) show the true effect size of stereotype threat implemented in the population on the x axis, and the variance in the population is given on the y axis. The plotted values range from 0 - 0.5, reflected in the colour gradient from green to white (see legend), and reflects the estimate that each analysis has produced in the model, averaged over the 20 repeat simulations of 60 datasets.

All analysis types in these plots are accurately estimating the size of the true effect, as can be seen by the colour gradient passing from green to orange along the x axis, as the true effect in the population increases, and the estimates from the analysis increase to reflect this. For example, when the true effect is zero, the plotted value is also equal to zero for all levels of variance. All analysis types are performing similarly, however as the baseline variance increases, the anova estimates shift away from the true effect size and slightly underestimate the true effect, as can be seen by the strips of colour bending slightly to the right as the variance increases. All other analysis types are accurately estimating the effect size regardless of variance.

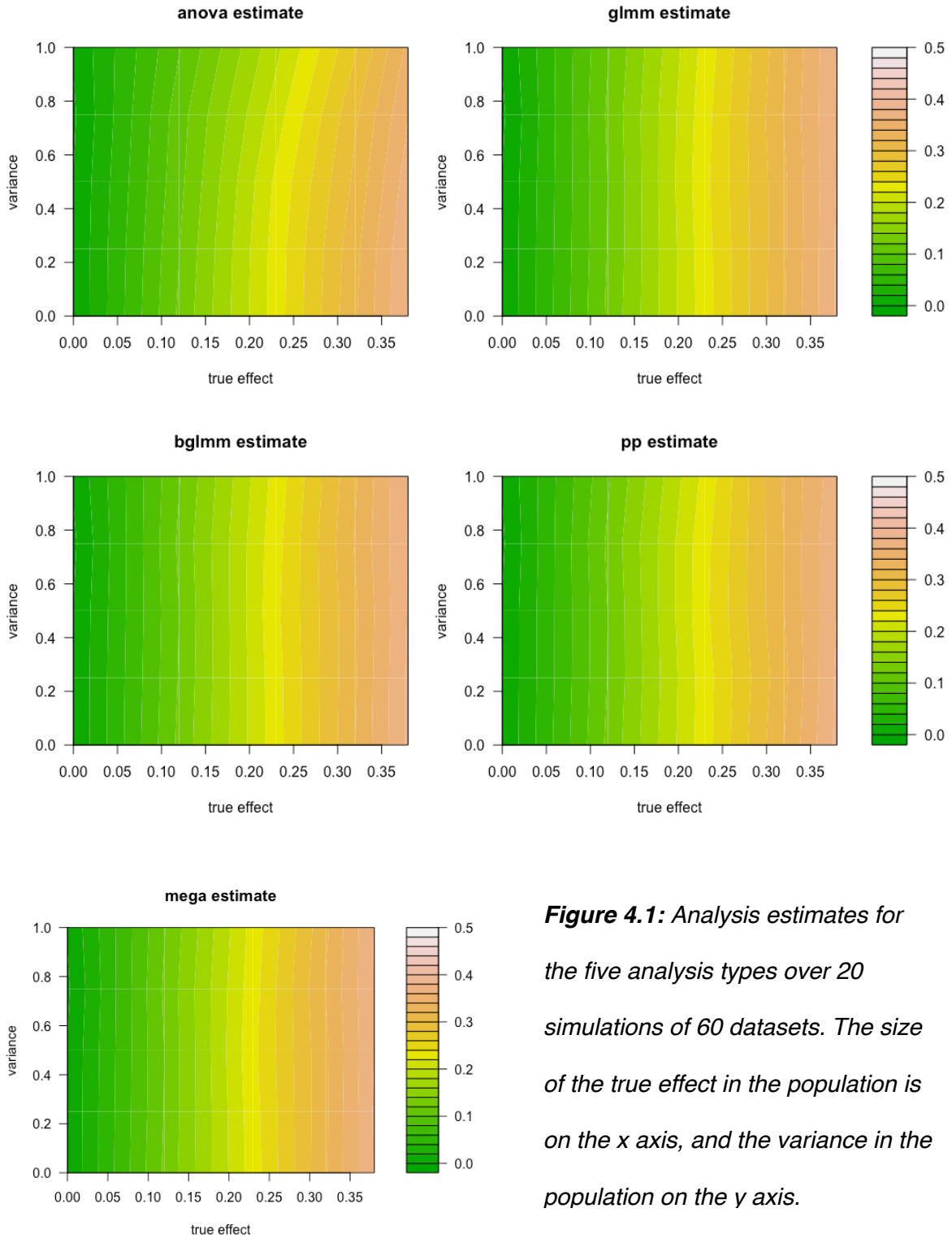


Figure 4.1: *Analysis estimates for the five analysis types over 20 simulations of 60 datasets. The size of the true effect in the population is on the x axis, and the variance in the population on the y axis.*

4.3.2 Positive Rate

The following plots (**Fig.4.2)** show the rate at which each analysis type finds an effect. The true effect of the stereotype threat effect in the population is on the x axis, the variance in the population is on the y axis. The plotted value is the Positive Rate, ranging from 0 - 1, represented as graded colours from green to white (see legend). An analysis finds a positive effect in the population if its upper and lower estimates do not include zero, the proportion of analyses in which a positive effect is found is the Positive Rate. Positive Rate is calculated for each analysis type averaged over 20 repeat simulations of 60 datasets.

The plots show that the anova, glmm and bglmm are consistently finding a true effect after the true effect in the population reaches a size of around 0.2. As the variance around the effect increases, the anova, glmm and bglmm are less consistent in finding an effect. Once the true effect size reaches around 0.3, the anova, glmm and bglmm are consistently reporting an effect as being present. In contrast, both the *posterior passing* and mega bglmm are consistently finding an effect once the true effect size is just over 0.1. In contrast to the other analysis types, the amount of variance around the true effect does not impact upon the likelihood of the analysis finding the true effect.

When the true population average was 0 (i.e. there is no real effect), all analyses were relatively unlikely to produce positive results. However, such false positives did occur in some instances in all analyses. Out of 100 instances of populations in which the true effect size was zero (i.e. 20 repeats of 5 different variance levels), posterior passing only culminated in two false positives in its final analysis. However, one of these false positives

co-occurred with a false positive (and the only false positive) in the mega bGLMM also, indicating that this was a genuine, spurious effect in that particular sample. In the other analyses, false positives were produced in at least one out of the 60 datasets in 91/100 ANOVA simulations, 93/100 GLMM simulations and 90/100 bGLMM simulations. Thus, using posterior passing dramatically reduces the number of false positives produced over 60 datasets from a population in which the true effect size is zero.

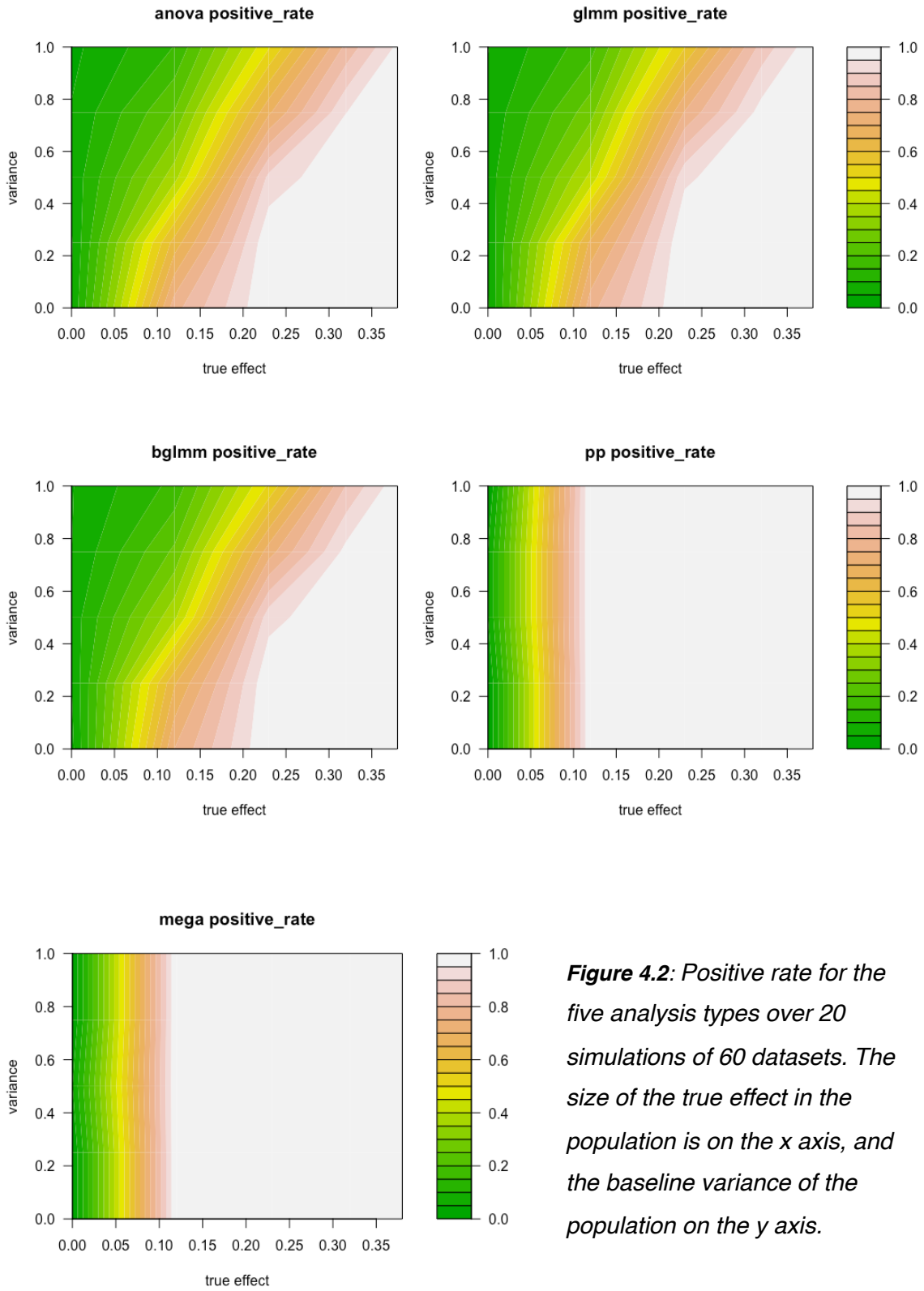


***Figure 4.2**: Positive rate for the five analysis types over 20 simulations of 60 datasets. The size of the true effect in the population is on the x axis, and the baseline variance of the population on the y axis.*

4.3.3 Uncertainty around the estimate

The following plots (**Fig. 4.3)** show the level of uncertainty of each analysis type. This measure is calculated as the size of the 95% confidence intervals, by subtracting the lower estimate from the upper estimate. The x axis shows the true effect in the population, and the y axis shows the variance in the population. The plotted value represents the Uncertainty level of each analysis, and ranges from 0 - 0.5, represented as a graded colour from green to white (see legend).

The anova plot shows that as the variance around the baseline increases, the uncertainty around the estimate also increases. As the true effect in the population increases, the uncertainty around the effect decreases slightly. The glmm and bglmm plots show a similar pattern to the anova plot, in that the uncertainty around the estimate increases as variance increases. The glmm and bglmm are less uncertain as the variance increases and as the true effect size increases. In contrast, both the *posterior passing* and mega bglmm plots show that these analysis types have very low uncertainty around their estimates, across all true effect sizes and all levels of variance, as all areas of the plot are green, reflecting almost zero uncertainty.

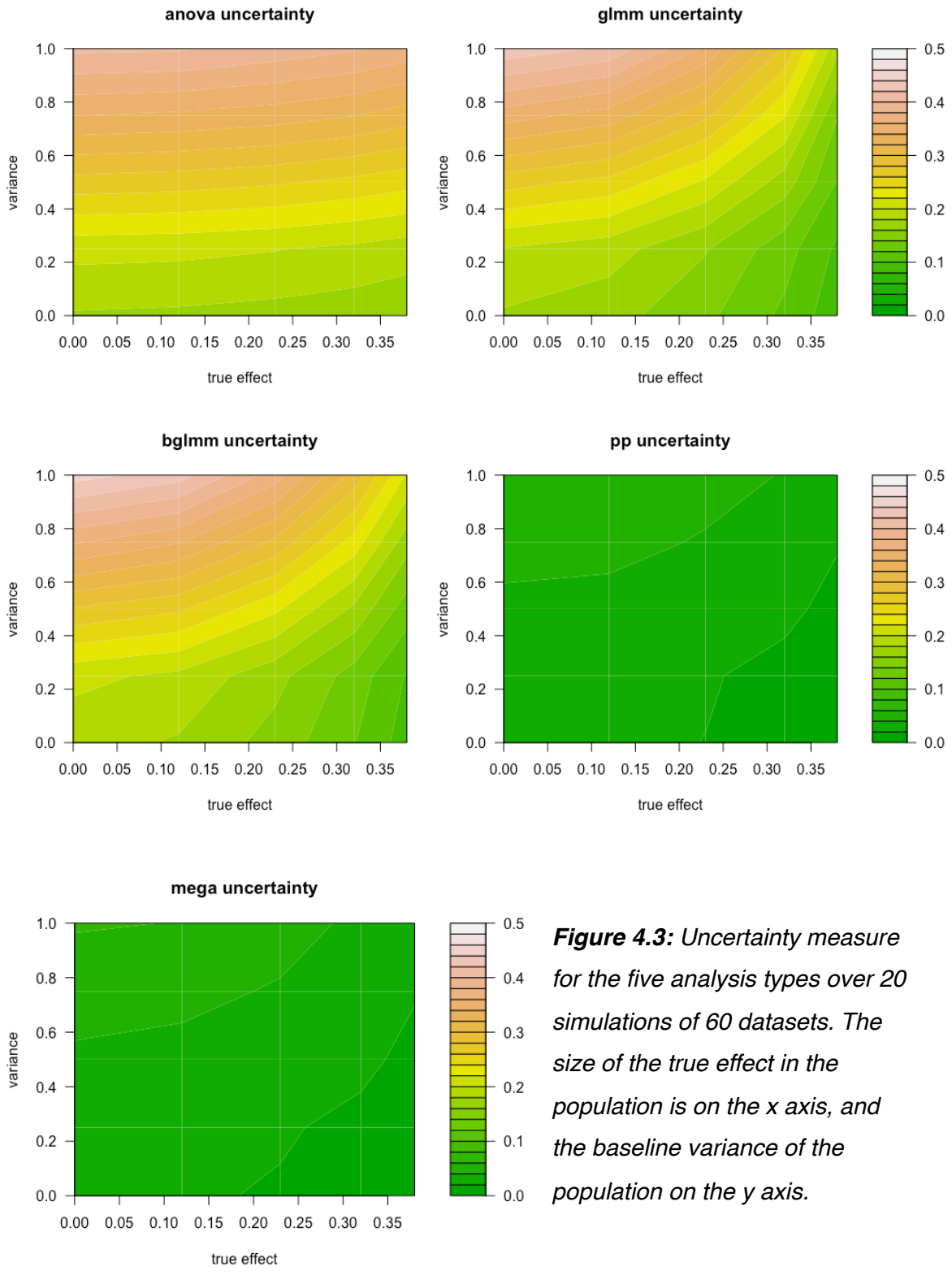


Figure 4.3: *Uncertainty measure for the five analysis types over 20 simulations of 60 datasets. The size of the true effect in the population is on the x axis, and the baseline variance of the population on the y axis.*

4.3.4 Error

The following plots (**Fig.4.4**) show the error around the estimate for each analysis. This measure is calculated as the difference between the estimate of the effect and the true effect in the population. The x axis shows the true estimate in the population and the y axis shows the variance in the population. The plotted value is the Error in each analysis, ranging from -0.1 to 0.1, represented as a graded colour from green to white (see legend).

The anova plot shows that the estimate has very little error when the true effect is small and the variance is low, but as the variance increases and the true effect increases, the estimate is prone to higher levels of error around the estimate, as the plot becomes more green in the top right quadrant of the plot. In contrast, all other analysis types show very little error, regardless of the true effect or level of variance in the population. This is reflected in all other plots consisting of only orange and yellow colouring, reflecting ranges close to zero levels of error.

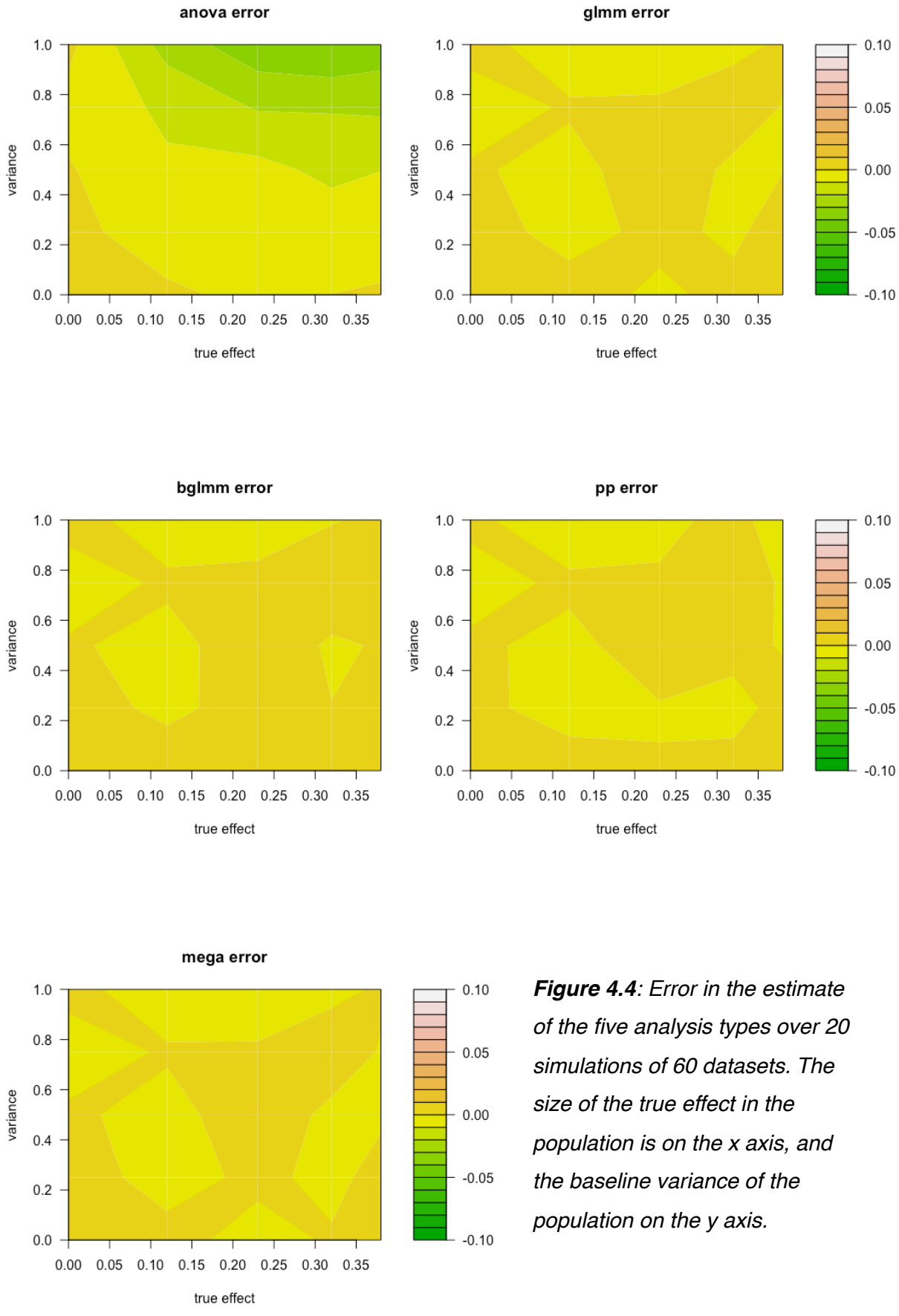


***Figure 4.4**: Error in the estimate of the five analysis types over 20 simulations of 60 datasets. The size of the true effect in the population is on the x axis, and the baseline variance of the population on the y axis.*

4.3.5 PP v. mega bGLMM

As a means to evaluate the success of our particular implementation of posterior passing (the passing of the mean and variance of the posterior distribution of the interaction effect), we included the 'mega bGLMM' in which all raw data from the 60 datasets were combined into a single dataset, and one single Bayesian GLMM was performed. Ideally PP would be as effective as the "mega bGLMM". The results support this with the results of PP and the mega bGLMM being almost indistinguishable and PP resembling the mega bGLMM to a much greater extent than any of the other analyses managed to (see Figs. 1-4).

4.3.6 Cumulative estimates over sequential datasets

The following plots (**Fig.4.5, 4.6**) are an illustrative example of the results using one simulation over 60 datasets when the true effect of the population is 0 and 0.38 respectively (shown by the horizontal black line on each plot). 0.38 was chosen as it is the largest effect size that we simulated in the population, given most meta-analyses report effect sizes of around 0.2 in the stereotype threat literature (e.g. Doyle & Voyer 2016). The variance parameter was held constant at zero throughout each simulation. The x axis represents each data set, numbered from 1 to 60. The y axis is the estimate that each analysis finds in each dataset, and the black horizontal line gives the true effect size in the population. Error bars represent 95% confidence intervals.

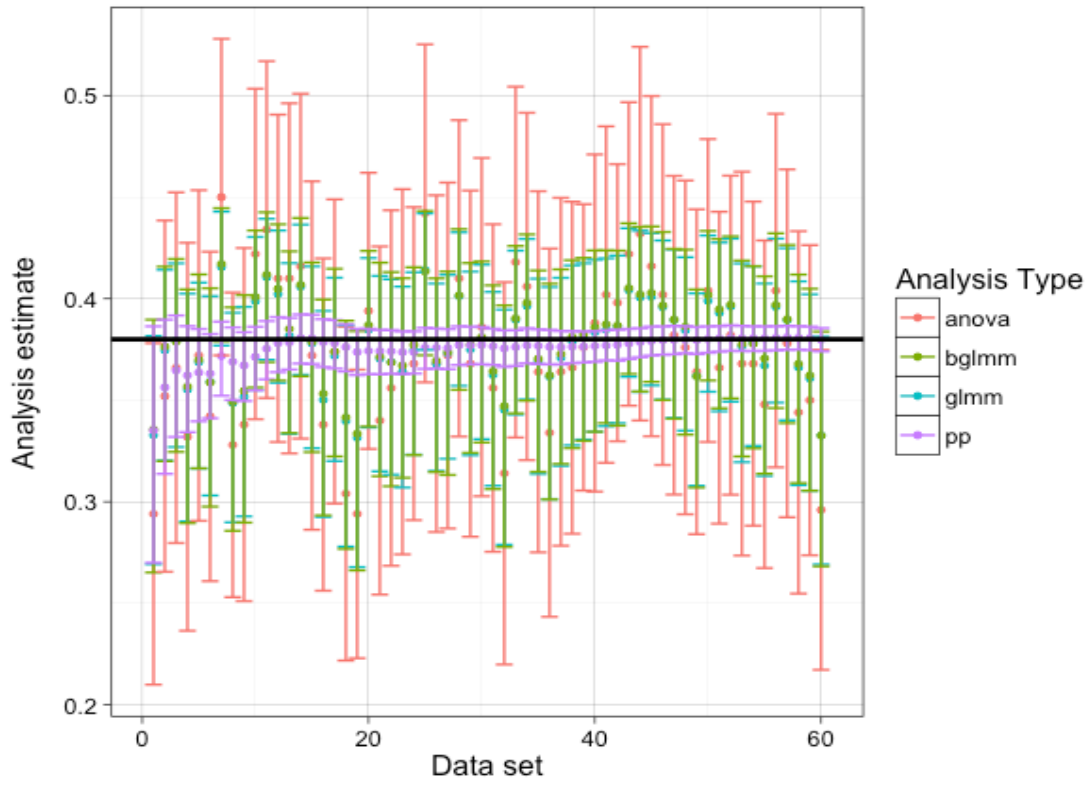


***Figure 4.5**: Analysis estimates and confidence intervals for one simulation of 60 datasets, with each data set on the x axis, and the analysis estimate on the y axis. The true estimate in the population is 0.38, displayed by the horizontal black line.*

Figure 4.5 shows each analyses' estimates of each data set when the true effect size in the population is 0.38 and the baseline variance is 0. The estimates for anova, glmm and bglmm vary around the true estimate, sometimes overestimating and sometimes underestimating the effect in the population. In stark contrast, posterior passing allows the analysis to become more certain over time, as each analysis uses information from previous analyses. Therefore over time (as each sequential dataset is analysed), posterior passing not only estimates the true effect size consistently, but becomes more and more certain about the presence and size of the effect, reflected by the narrowing of the confidence intervals. Furthermore, the anova estimates are considerably less certain than the glmm and bglmm estimates, as their error bars are much wider across all the datasets.

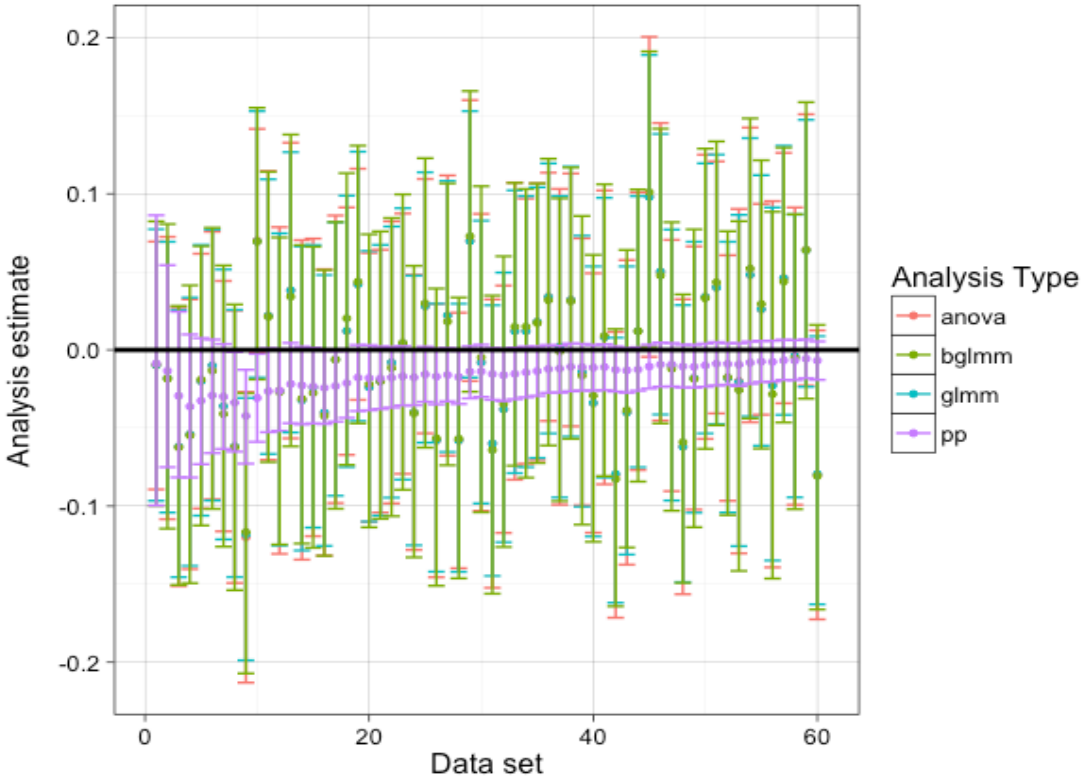


Figure 4.6 *Analysis estimates and confidence intervals for one simulation of 60 datasets, with each data set on the x axis, and the analysis estimate on the y axis. The true estimate in the population is 0, displayed by the horizontal black line.*

Figure 4.6 shows each analyses' estimates of each data set when the true effect size in the population is 0 (ie when there is no stereotype threat effect present in the population). The anova, glmm and bglmm analyses all vary around the true estimate, sometimes estimating a small positive or negative effect when there is none, although the confidence intervals mostly encompass zero. In contrast, posterior passing produces estimates close to zero from early on in the sequence, and gradually get more certain as more data is encompassed into each analysis. Interestingly, at data set 9, all analyses estimate a negative effect, presumably due to an unusual sample drawn from the population, however, posterior passing is not dramatically affected by this dataset and soon converges on the true effect again.

4.4 Discussion

I simulated 60 datasets of varying effect sizes and variance to represent the stereotype threat literature from a simulated population of one million individuals. I analysed all data sets using four different analysis methods; two frequentist and two bayesian methods. I found that using a method of posterior passing to analyse many data sets allows the true effect in the population to be found with very high certainty, whereas conducting many separate analyses over the same data sets leads to uncertainty around the effect estimates. Posterior passing also consistently identifies the presence of an effect, as well as the size of the effect. On all effect sizes above 0.1, even with very high levels of variance, posterior passing accurately identifies the effect, whereas all other analyses are less consistent at identifying such small effect sizes, especially with higher levels of variance. Furthermore, posterior passing produces estimates as accurate and precise as the 'mega bglmm' comparison, which conducts one analysis on all of the data sets at once.

These results suggest that when posterior passing is used, there is considerably more certainty provided regarding the presence or absence of an effect in the population, regardless of the size of the effect or the variance around the effect. Not only does posterior passing provide more certainty on the presence or absence of an effect, but also the size of the effect, as the analysis becomes more certain over time. This can be clearly seen in the examples in **Figure 4.5, 4.6.** Furthermore, posterior passing is just as accurate as if all data had been analysed at once in one analysis, as can be seen when comparing the mega bglmm and posterior passing plots in **Figure 4.2**.

Overall, it can be seen that using a bayesian method of data analysis with informative priors would provide more certainty about the presence or absence of an effect in the population, when confronted with lots of conflicting datasets, such as is the case in the stereotype threat literature. In **Figure 4.5**, the orange points (anova) are analogous to the current culture of data analysis in the stereotype threat literature; anovas are the most common analysis

method. In this way, each dataset is analysed using null hypothesis testing, treating each data set as an isolated, one-off hypothesis test. If we view our 60 example data sets as representing years' worth of experiments, we can see how the anovas produce varying estimates of effect sizes with little certainty around the estimate. This uncertainty would potentially instigate several meta-analyses to attempt to get a clearer conclusion of the literature, as has been the case in the stereotype threat literature (e.g. see Flore & Wilcherts 2013). The purple points represent an accumulation of certainty over time as more data is collected- something which is claimed to be one of the strengths of the scientific process in general. Moreover, these two example simulations had a baseline variance of zero, which, based on the results of all simulations, all analyses should be performing best at. Therefore these examples are, if anything, the best case scenario for the alternative methods.

It has previously been argued that bayesian inference solves the multiple comparisons problem that occurs when conducting multiple statistical tests (Gelman & Tuerlinckx 2000), as Bayesian inference is more conservative in that it goes with what is already known unless the data forces a change (Gelman 2010). Additionally, Bayesian methods with informative priors improve inference as more information is used, so better estimates of effect sizes are produced (Gelman 2010). In a similar vein, Beppu & Griffiths (2009) found that a posterior passing approach led to successively better inferences over time in a lab experiment and a theoretical model of simulated participants. Indeed, information about the certainty and plausibility of information, as well as the observations themselves, are needed for knowledge and inference to cumulatively improve over time (Beppu & Griffiths 2009). Their findings run parallel to the results of these simulations, in which each analysis uses not only new data, but updates its "prior beliefs" with the inference from the previous data set. This approach is a more intuitive, efficient and ultimately more accurate way for science to proceed, in that all data are used to update beliefs about the world.

An incentive for all data to be incorporated into analyses would not only encourage null results and negative findings to be published, but also collaboration rather than competition between labs, as all data is equally valuable in helping to improve inference. Indeed as the number of journals and papers in scientific publishing are increasing, more and more independent labs are exploring the same effects. With an incentive to combine data sets and incorporate all information for better inference, rather than multiple papers with conflicting conclusions, disparate research findings could be combined into collaborative publications. Indeed, posterior passing may only be effective if all data is incorporated; if findings are "file-drawered" posterior passing may be over confident in the presence of an effect as the negative results are not incorporated. A "file-drawer" effect could be explored with this model in future work. For example, by implementing varying levels of publication bias we could explore to what extent the number of false positives present in the literature is exacerbated or alleviated by a posterior passing approach.

Along with discouraging "file-drawing" experimental results, the use of posterior passing could also encourage the implementation of pre-registered methods. Indeed, if all datasets are to be published, then pre-registration is necessary to prevent "HARKing", which is a problem tied up with publication bias. If datasets are only published in the light of supported predictions and positive results, then the contradictory and confusing mass of literature, such as the stereotype threat literature, is inevitable as researchers justify their unexpected results with post-hoc, alternative theories to those originally being investigated. However, if researchers were to pre-register their theories, methods and predictions in order to contribute whatever their findings to the cumulative nature of a posterior passing approach, the problem of harking is eliminated.

If the nature of publishing was so that all available data can be used towards a cumulative analysis, hopefully this will dramatically reduce the number of wasted research hours, funds and efforts that go towards studying a "null" effect that appears to be present in the literature

due to publication bias, deliberate or unintentional p-hacking, or purely due to the inevitability of finding a “significant” effect after many repeated null hypothesis tests. Additionally, a move away from null hypothesis testing and blunt uses of statistics to detect “a difference” may encourage a more nuanced interpretation and reporting of results, such as presenting individual variation and overlapping frequency distributions. This has been called for particularly in the reporting of sex differences by Maney (2015), who calls for statistics that explain the extent to which sex explains variation, rather than a “yes or no” as to whether sexes differ. This could help to reduce common misinterpretations and oversimplifications that often catch the attention of the media and become miscommunicated to the public, further increasing erroneous perceptions of sex differences that are not supported by the research results.

However, the benefits of posterior passing are not without caveats and before endorsing a transition to the use of informative priors it is worth noting the possible limitations of the proposed method. In the current simulations, we created a set of identical, sequential datasets. However, in reality, scientific experiments are rarely identical, nor do they happen in regular sequence with one experiment neatly following on from another. Therefore, one question that arises for a “real world” implementation of informative priors is where to get the relevant data to form the priors. One answer would be to derive the prior from an existing dataset, or group of datasets, most relevant to the research question, containing an effect most similar to the effect of interest. Of course, these judgements will inevitably be subjective, but the uncertainty could be reflected in the strength of the priors that are implemented. For example, the strength of the prior could vary depending on the size of the dataset in question and the number of previous relevant datasets available; researchers could vary the prior to give an upper and lower range of the estimate in question. Appropriate use of priors will be an important issue for researchers, and until norms are decided upon it would be necessary for researchers to make clear which priors were used and why. Indeed, a constructive means of reviewing a paper could be to explore weaker and

stronger priors to test the robustness of the reported effect. Although concerns are often raised around the nature of priors “nudging” analyses into a desired direction, our simulations show that posterior passing did not show signs of distortion, even after a spurious posistive result (visible in **Fig. 4.6**, dataset 9). However, further explorations and implentations of informative priors would be valuable to test the robustness of posterior passing in different scenarios. Such work would be valuable for providing an evidential basis before adopting new research practices.

This model has revealed a benefit to using informative priors when many labs are sequentially and simultaneously studying the same effect in the population. This approach to data analysis could be a potential step forward for the field of psychology in the light of the replication crisis, as well as being a closer approximation to the scientific process of accumulating knowledge over time to improve inference. I hope that this model will provide a starting point for exploring many hypothetical scenarios to which posterior passing may be beneficial. In future research this model can be used to explore how posterior passing performs with varying sample sizes, trial numbers, non-linear data, spurious results and the presence of publication bias.

Chapter 5: Discussion

Discussion

5.1 Summary of findings

In **Chapter 2,** I used a novel computer-based task to explore whether varying the pay-offs of social and asocial sources of information differentially affected men and women's tendency to choose those sources. I found that when asocial information was manipulated to be riskier than social information, women showed a preference for social rather than asocial learning compared to when neither asocial nor social information was risky. In contrast, men showed a preference for asocial learning when neither were risky, as well as when asocial learning was the riskier option. Both men and women also showed a preference for asocial learning in the condition in which social learning was the riskier option. When comparing the two conditions in which the riskiness of the options differed, both men and women were more likely to choose the risky option when given a low rank. I also found that men scored higher, on average, on a risk-taking measure than women did, supporting previous findings that men score higher on risk-taking measures than women do (Campbell & Muncer 2009). The risk-taking scores did not predict the tendency of an individual to choose social or asocial learning, nor did it predict whether individuals chose the risky or safe option; however this was not the main hypothesis being tested in this study.

In **Chapter 3,** I attempted to manipulate men and women's confidence on a mental rotation task using stereotype threat primes, as well as investigating whether altering men and women's confidence affected their likelihood to conform to a majority's answer on the task. In Study 1, I found that men and women both gave higher confidence ratings when primed to believe that women perform better on the task, compared to men and women in the control

condition. However, this result was not replicated in the subsequent two studies. In Study 2, men and women's performance, but not confidence, was higher on average when primed to believe that women perform better on the task. In Study 3, neither men nor women's confidence, or performance, was affected by any of the stereotype threat primes. After seeing social information in Study 2 and 3, participants were more likely to switch their answer when their confidence ratings were low compared to when their confidence ratings were high. Furthermore, in both Study 2 and 3, women gave lower confidence ratings on average than men, and this sex difference in confidence had an indirect effect on switching, as lower levels of confidence led to higher levels of switching.

In **Chapter 4,** I simulated multiple datasets based on the stereotype threat literature and analysed them using four different methods of analysis. I found that using Bayesian analysis methods with informative priors in a posterior passing approach led to higher certainty about the presence or absence of an effect in the population over many cumulative analyses than the other analysis methods. Furthermore, the precision and size of the estimate was more certain than with other forms of analysis such as frequentist GLMM's or ANOVAs. I found that the posterior passing approach was just as accurate at estimating the effects as a 'mega GLMM' in which all datasets were analysed as one large dataset. Moreover, ANOVAs, which represent the most commonly used method to analyse psychological priming studies such as stereotype threat, performed most poorly compared to the other analysis methods, especially when faced with a population with high variance.

5.2 Implications of sex differences in social and asocial learning

The finding from **Chapter 2**, that women preferentially chose social sources of information when asocial sources were riskier in their pay-offs, whereas they preferentially chose asocial learning when both sources were equal in their pay-offs, supports previous bodies of work suggesting that social learning is used strategically and flexibly (Laland 2004, Hoppitt & Laland 2014). It supports the theory that using social or asocial learning consists of a trade-off between the costs and benefits of each (Kendal et al. 2005), in that women behaved differently when the sources of information differed in their expected pay-offs to when the pay-offs were equal. The finding that men and women differed in their choices to learn socially or asocially when asocial information was riskier than social, presumably due to men's higher risk-taking preferences, supports literature suggesting that personality differences such as boldness and risk-taking influence the likelihood to learn asocially or socially (see Brosnan & Hopper 2014), and that individual differences in social and asocial learning are worth further investigation (Mesoudi et al. 2016). **Chapter 2** also provides one of the first studies to explicitly investigate whether risk-taking differences in adult humans are linked to sex differences in the preference to use social or asocial sources of information.

Although sex differences in behaviour are often discussed in terms of sex differences in the brain (see Prager 2017; Cahill 2017), sex differences in the context of social learning are rarely discussed. When sex difference research is reduced to anatomical or physiological differences visible in the human brain, researchers (as well as the media coverage of the research) often infer that such differences stem from fixed, genetic, or evolutionary origins (e.g. Ingalhalikar et al. 2014). Although physical differences in the brain are in many respects easier to observe and measure than differences in behaviour or cognition, the presence of such differences does not give any insight into whether those differences arose due to

evolutionary reasons, developmental processes, cultural evolution, social factors, or most likely, a mixture of all of these variables. In contrast, studying sex differences in social learning within a cultural evolution framework induces an awareness of the complex interactions and feedback loops between all of these factors, as well as providing opportunity to manipulate some of these factors to explore the plasticity of the behavioural outcomes. The topic of sex differences in human behaviour provides an excellent avenue to study these complex processes and, perhaps due to controversy within academia and the media (e.g. Sample 2013; Wheeling 2015), forces researchers to question whether behavioural outcomes can be reduced to purely biological or cultural factors. I believe both are always necessary to gain a full understanding of how sex differences arise, as with how any behaviour arises. With the same token, it is unlikely that sex is in and of itself an explanatory variable for behavioural differences. An individual's sex can be viewed as a proxy for other factors that are yet to be uncovered, and investigating the extent to which sex explains variation in behaviour, rather than simply whether the sexes differ or not, is crucial (Maney 2016).

As reviewed in **Chapter 2**, individual and personality differences such as risk-taking tendencies can affect the tendency to use social or asocial information. However, these risk-taking tendencies may also be shaped by an individual's social or cultural environment itself (see Nelson 2015). Physiological factors (e.g. differences in hormone levels), and evolutionary explanations (e.g. differences in competition or parental investment) can account for one sex being more risk-taking than the other (e.g. Apicella et al. 2008; Wilson & Daly 1985; Campbell 1999). However, the cultural environment can also put different pressures on men and women; potentially reinforcing, ameliorating or exacerbating any pre-existing differences in risk-taking preferences (see Nelson 2015). Both of these factors can then affect how, if and when individuals socially learn. If men and women's cultural environment has an indirect effect on how likely they are to learn socially or asocially,

mediated by risk-taking differences, these differences could have a reinforcing effect on resulting sex differences. For example, if women are encouraged to be less risk-taking by their social or developmental environment, this could indirectly cause them to choose social rather than asocial information sources, further reinforcing their lower risk-taking preferences and further increasing their likelihood to choose socially, thus exaggerating the sex difference in risk-taking and social learning even further. The process by which men and women are differentially affected by social learning, which in turn leads them to use social information more or less than members of the other sex, is of great value to social learning researchers. Incorporating potential sex differences into theoretical models of how social information is transmitted could have important consequences for predictions of population level processes of social learning. For example, these population-level models could be used to explore the broader patterns and effects on real world examples of behaviour, such as occupation choice, creative activities, leisure activities, health choices and political decisions.

5.3 Implications of sex differences in confidence and conformity

Just as personality and individual differences in risk-taking behaviour are likely to affect men and women's propensity to use social or asocial learning, **Chapter 3** revealed that differences in confidence also impact upon men and women's decision to conform to a majority decision, in that lower confidence led to an increased likelihood of switching answer after seeing social information. This finding supports previous work showing that lower confidence leads to higher likelihood of conforming, suggesting that people use a copy-when-uncertain social learning strategy (Morgan et al. 2012). It also further supports the work of Cross and colleagues (2016) who found that lower average confidence in women

compared to men on a mental rotation task had an indirect effect on conformity. Thus, previous findings that women conform more than men (e.g. Bond & Smith 1996) may be better understood as a downstream effect of sex difference in confidence.

Men and women's different confidence levels may be affected by physiological factors such as differences in testosterone levels (Johnson, Zava & McCoy 2000) and may also be linked to individual and personality differences in traits such as neuroticism, extraversion or assertiveness (Costa et al. 2001; Judge et al. 2002; Schaefer et al 2004). However, it has been found that confidence may not only be "a state of being certain about the success of a particular behavioural act" (Stankov et al. 2012) but also a structurally distinct psychological trait, lying somewhere between cognitive ability and personality (e.g. Blais et al. 2005; Stankov & Lee 2008). Confidence relies on an internal assessment of how likely you are to be correct, or successful, before or after performing a particular behaviour. Accordingly, there is evidence for overconfidence and underconfidence in a range of tasks (e.g. Johnson & Fowler 2011), where an individual's ability to accurately assess their success is impaired. Thus, receiving feedback on performance contributes significantly to an individual's self-confidence. Furthermore, feedback on success relative to other individual's success will contribute to an individual's self-confidence, such as explicit rankings within a laboratory task or within real world scenarios. In this way, women's differential treatment to men in the workplace, such as lower pay-rates (Arulampalam, Booth & Bryan 2007), lower likelihood of being hired (Moss-Racusin et al. 2012), lower likelihood of being invited to talk at a conference (Shroeder et al. 2013) qualitatively different reference letters (Trix & Psenka 2003; Schmader, Whitehead & Wysocki 2007), and lower likelihood of employment in STEM areas (Blickenstaff 2005) may directly be impacting upon women's confidence differently to men's, which in turn may exacerbate their likelihood to conform in these environments.

An individual's confidence may also interact with other traits such as dominance and prestige, which influence patterns of social learning through prestige-biased or dominance-

biased social learning. Furthermore, how men and women appraise themselves in comparison to one another may also have consequences for how likely men and women are to become prestigious or dominant in different contexts (Cheng et al. 2013). Theories on dominance and prestige suggest that individuals may become prestigious if other individuals recognise their success in a specific domain, consequently conferring deference to these individuals (Henrich & Gil-White 2001). Convincing others that you are successful in a particular domain will not only rely on one's success but also portraying and communicating that success to others. Portraying one's success to others convincingly relies on a certain level of self-confidence. In the same vein, behaving in a dominant way such that other individuals are influenced by one's behaviour relies upon confidently asserting oneself relative to others. Both dominance and prestige have been found to impact upon how social information is transmitted and how social hierarchies are formed (Cheng et al 2013; Henrich & Gil-White 2001). Thus confidence, and sex differences in confidence, are likely to interact with whether individuals become prestigious or dominant in a social hierarchy, which can in turn affect men and women's social information use. For example, if a context arises in which women are disproportionately at the bottom of a social hierarchy and are gaining negative feedback relative to men, this may lead to a lowering of confidence in women relative to men, increasing women's likelihood of conforming to others, and reducing their likelihood of becoming prestigious or dominant enough to be copied by others. Again, a better understanding of how factors such as dominance, prestige and confidence interact is necessary to fully understand cultural micro and macro-evolutionary dynamics (Mesoudi et al 2016).

5.4 Implications of Bayesian analysis methods in psychology

In **Chapter 4,** I found that using a Bayesian method of analysis that involves informative priors and posterior passing leads to greater certainty about the presence or absence of an effect in a population than the other statistical methods. This finding has important implications for the way scientists conduct, analyse and publish their research. Firstly, the current norm for analysing psychological priming studies, such as stereotype threat, is the ANOVA. In my simulations of multiple datasets with varying effect sizes and variation in the population, the ANOVA was the least accurate at identifying the effect size, especially when the effect size was small and the variation high. The priming literature is precisely when researchers predict effect sizes to be small and variance to be high, as the mechanism underlying the effect is unknown and some individuals are expected to be more or less susceptible to the effect depending on various moderating variables (see Bargh 2012; Gelman 2016). Therefore, the fact that ANOVAs are less likely than other methods to be able to accurately decipher an effect in these types of datasets suggests that researchers should move away from the ANOVA on to other methods, as has been previously suggested (e.g. Jaeger 2008).

I also found that the Bayesian GLMM with minimally informative priors did not provide any detectable improvement in estimates compared to the frequentist GLMM. This was expected, as the two methods are almost identical, both using maximum likelihood estimation to find the most likely estimate, and minimally informative priors used in the Bayesian version are not strong enough to override the patterns present in the data. Thus, there is little inherent benefit to researchers moving to Bayesian methods of analysis over frequentist methods, if priors are habitually implemented uninformatively. One benefit would be the move away from null hypothesis significance testing and towards the more intuitive philosophy of Bayesian inference (McElreath 2016), as well as having access to the full

posterior distribution of parameter estimates, rather than a single point estimate. However, for gaining certainty over whether an effect is absent of present in a population given many datasets, using a Bayesian GLMM with minimally informative priors would not improve certainty compared to a frequentist GLMM. In contrast, using a posterior passing approach which implements informative priors based on the estimates of previous analyses greatly improves certainty and accuracy of the estimates of effects in a population.

The use of posterior passing has implications for scientists who are all presumably researching the same underlying phenomenon or effect, such as the stereotype threat effect. However, posterior passing clearly will only be an improvement if all scientists publish all of their research, regardless of whether they find evidence for an effect or not. Indeed, posterior passing may exacerbate the problem of publication bias if researchers only put forward their positive results to be part of a cumulative posterior passing framework. Thus, researchers need to be incentivised to publish all findings towards a cumulative, collaborative investigation into the same effect, in which it is recognised that all evidence about the absence or presence of the effect is valuable and necessary to uncover the most accurate and reliable estimates of the effect (Smaldino & McElreath 2016). One way in which this practice may be incentivised is for pre-registration to become compulsory, in that researchers publish their planned methods and analyses before data are collected, on the assurance that whatever their results, their work will be published. This framework of pre-registering methods and publishing all findings to be included in a cumulative analysis of all evidence would greatly help to eradicate false positives and uncertainty within scientific literature.

If these recommendations are put into practice, researchers could not only be more certain about the reliability of the literature they read and cite to inform their work, but this certainty would also arise more rapidly and efficiently. As more data are incorporated into a cumulative analysis, each analysis will have information from all previous analyses. For

example, in a hypothetical scenario of many studies examining the same effect, study number one will have no choice but to incorporate minimally informative priors as there will be no previous data to use at this point. Study number two could then use a weakly informative prior using the estimates from Study number one, but with large amounts of uncertainty in the prior so that the estimate is derived predominantly from the data. This process could continue so that with each new dataset, the analysis uses more information and stronger priors. Accordingly, fewer studies will be needed overall to gain greater certainty in an effect being present or absent in a population, as each dataset will not be analysed independently from all others, but in the context of the rest of the data available. As was shown in **Chapter 4**, this posterior passing technique rapidly stabilises on the correct estimate, thus rendering further studies, research time, effort and resources as unnecessary.

5.5. Future work

Future work exploring the factors that lead to sex differences in risk-taking would be useful for predicting in what contexts these sex differences are likely to lead to differential use of social and asocial information. One way to achieve a greater understanding of the factors influencing men and women's risk-taking preferences would be to gather more cross-cultural data of a range of risk-taking measures to decipher which social and cultural environments lead to larger or smaller sex differences in risk-taking. In a similar vein, altering the riskiness of social and asocial information using a continuum of pay-off differences would allow researchers to decipher the cut-off point at which men and women are dissuaded from using these information sources. Finding a task in which participants have no baseline preference for social or asocial information sources would allow researchers to explore which factors lead men and women to choose social or asocial sources of information differently depending on pay-off differences. This type of task would ideally be used with a range of participants from different social and cultural environments, with a range of risk-taking

preferences. Furthermore, larger sample sizes would allow finer-grained exploration of how individual differences in risk-taking measures might correlate with social and asocial learning preferences.

Another avenue of further investigation would be to use different methods of manipulating men and women's confidence, perhaps by using false feedback on other tasks. This approach could be a more reliable method of artificially altering confidence in an experimental setting than stereotype threat primes. Additionally, collecting men and women's self confidence ratings in different domains, and in response to real world feedback such as job references, interview feedback, promotions or invited talks would be valuable data for understanding how one's feedback relative to others affects confidence, which in turn affects social information use. Understanding how personality differences, feedback and self-confidence interact with the likelihood of becoming a prestigious or dominant individual in a social hierarchy is something that I hope to explore in the future.

Finally, when modelling the effect of using different analysis methods on a body of literature, a file-drawer effect could be implemented to explore the impact of only publishing positive results on the use of posterior passing. The model presented in **Chapter 4** could be adjusted to include a file-drawer effect in which only positive results are incorporated in the final analyses. The model in **Chapter 4** could also be used to explore other scenarios in scientific research. For example varying the sample sizes in the simulated data, or varying the trials per participant, could produce datasets that reflect fields of research other than the stereotype threat literature. Furthermore, implementing the use of informative priors in the analysis of real world datasets, for example among collaborating labs who are investigating similar effects, would be valuable in demonstrating how this type of analysis would impact upon researchers' inferences and publications. One possibility would be to analyse data both using strongly informative and weakly informative priors, to give an idea of how conclusions

differ when analysing datasets in isolation compared to analysing them within the context of other similar datasets.

5.6 Conclusion

In this thesis, I set out to explore the contexts in which sex differences in social and asocial information use occur, and whether these differences can be understood in terms of other factors such as risk-taking and confidence. I found that when asocial learning was manipulated to be riskier, women were more likely to choose to learn socially compared to women in the control condition. This pattern did not emerge in men. I also found that men and women differed in their confidence ratings in a mental rotation task, and this difference had an indirect effect on their likelihood of conforming to a majority's answer on the task. Further to this, I explored whether men and women's confidence could be affected by socially transmitted stereotypes, however the literature examining the effect of stereotypes on men and women's confidence and performance is conflicted and difficult to interpret. The stereotype threat literature is likely to have been subject to publication bias, which makes interpreting the effect of stereotypes difficult. I found that such a conflicting literature may be prevented by employing Bayesian methods of analysis with posterior passing, in order to decipher the true presence or absence of an effect in a population. In light of these findings, I conclude that sex needs to be addressed as a variable of interest, as something that is shaped by culture and, through the interaction of many contributing factors, also has an impact on cultural transmission itself. Having said that, sex should not be viewed as an explanatory variable in and of itself, but as a proxy that points to other processes that remain uncovered, such as individual differences in personality traits or in how men and women appraise their success and confidence in different contexts. Furthermore, the successful study of complex psychological and behavioural effects, such as those put forward in the

stereotype threat literature, relies upon scientists being rigorous, transparent and collaborative in their methods. To this end, scientists studying such complex behavioural phenomena would benefit by using pre-registered methods, publishing all results, and putting all results towards a collective, collaborative and cumulative analysis, so that the presence or absence of an effect can be cumulatively better understood as a whole.

References

Acerbi, A., Van Leeuwen, E. J., Haun, D. B., & Tennie, C. (2016). Conformity cannot be identified based on population-level signatures. *Scientific Reports*, *6*.

Alexander, G. M., & Evardone, M. (2008). Blocks and bodies: Sex differences in a novel version of the Mental Rotations Test. *Hormones and behavior,* 53(1), 177-184.

Apicella, C. L., & Dreber, A. (2014). Sex Differences in Competitiveness: Hunter-Gatherer Women and Girls Compete Less in Gender-Neutral and Male-Centric Tasks. *Adaptive Human Behavior and Physiology*, 1-23.

Aplin, L. M., Farine, D. R., Morand-Ferron, J., Cockburn, A., Thornton, A., & Sheldon, B. C. (2015). Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature*, *518*(7540), 538-541.

Arbilly, M., Motro, U., Feldman, M. W., & Lotem, A. (2011). Evolution of social learning when high expected payoffs are associated with high risk of failure. *Journal of the Royal Society Interface*, *8*(64), 1604-1615.

Archer, J. (2006). Testosterone and human aggression: an evaluation of the challenge hypothesis. *Neuroscience & Biobehavioral Reviews*, *30*(3), 319-345.

Arulampalam, W., Booth, A. L., & Bryan, M. L. (2007). Is there a glass ceiling over Europe? Exploring the gender pay gap across the wage distribution. Industrial & Labor Relations Review, 60(2), 163-186.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men*, 222-236.

Atkisson, C., O'Brien, M. J., & Mesoudi, A. (2012). Adult learners in a novel environment use prestige-biased social learning. *Evolutionary psychology*, *10*(3), 147470491201000309.

Bargh, J.A. (2012). *Priming Effects Replicate Just Fine, Thanks.* Psychology Today. from: www.psychologytoday.com

Barnard, C.J. and Sibly, R.M., 1981. Producers and scroungers: a general model and its application to captive flocks of house sparrows. *Animal behaviour*, *29*(2), pp.543-550.

Baron, R. S., Vandello, J. A., & Brunsman, B. (1996). The forgotten variable in conformity research: Impact of task importance on social influence. *Journal of Personality and Social Psychology*, *71*(5), 915.

Basow, S. A. (2010). Changes in psychology of women and psychology of gender textbooks (1975–2010). Sex Roles, 62, 151–152. doi:10.1007/s11199-010-9744-z.

Bateup, H. S., Booth, A., Shirtcliff, E. A., & Granger, D. A. (2002). Testosterone, cortisol, and women's competition. *Evolution and Human Behavior*, *23*(3), 181-192.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 1088-1101.

Bengston, S. E., Shin, M., & Dornhaus, A. (2016). Life-history strategy and behavioral type: risk-tolerance reflects growth rate and energy allocation in ant colonies. *Oikos*.

Beppu, A., & Griffiths, T. L. (2009). Iterated learning and the cultural ratchet. *In Proceedings of the 31st annual conference of the cognitive science society* (pp. 2089-2094). Austin, TX: Cognitive Science Society.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis*, *20*(3), 351-368.

Beyer, S., & Bowden, E. M. (1997). Gender Differences in Self-Perceptions: Convergent Evidence from Three Measures of Accuracy and Bias. Personality and Social Psychology Bulletin, 23(2), 157-172.

Biro, P. A., & Stamps, J. A. (2008). Are animal personality traits linked to life-history productivity?. *Trends in Ecology & Evolution*, *23*(7), 361-368.

Blais, A. R., Thompson, M. M., & Baranski, J. V. (2005). Individual differences in decision processing and confidence judgments in comparative judgment tasks: The role of cognitive styles. *Personality and Individual Differences*, *38*(7), 1701-1713.

Blanch, D. C., Hall, J. A., Roter, D. L., & Frankel, R. M. (2008). Medical student gender and issues of confidence. *Patient education and counseling*, *72*(3), 374-381.

Blickenstaff, J. C. (2005). Women and science careers: leaky pipeline or gender filter? *Gender & Education*, 17(4), 369 – 386.

Bohannon J. (2014) Replication effort provokes praise—and 'bullying' charges. *Science*. 2014; 344:788–789.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, *24*(3), 127-135.

Bond, R. (2005). Group size and conformity. *Group processes & intergroup relations*, *8*(4), 331-354.

Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological bulletin*, *119*(1), 111.

Boogert, N. J., Reader, S. M., & Laland, K. N. (2006). The relation between social rank, neophobia and individual learning in starlings. *Animal Behaviour*, *72*(6), 1229-1239.

Boose, K. J., White, F. J., & Meinelt, A. (2013). Sex differences in tool use acquisition in bonobos (Pan paniscus). *American journal of primatology*, *75*(9), 917-926.

Booth, A. L., & Nolen, P. (2012). Gender differences in risk behaviour: Does nurture matter?. *The Economic Journal*, *122*(558), F56-F78.

Bower, B. (2012). The hot and cold of priming: Psychologists are divided on whether unnoticed cues can influence behavior. *Science News*, 181(10), 26-29.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process.* University of Chicago Press.

Brosnan, S. F., & Hopper, L. M. (2014). Psychological limits on animal innovation. *Animal Behaviour*, *92*, 325-332.

Budaev, S. V., Zworykin, D. D., & Mochek, A. D. (1999). Consistency of individual differences in behaviour of the lion-headed cichlid, Steatocranus casuarius. *Behavioural Processes*, *48*(1), 49-55.

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological bulletin*, *125*(3), 367.

Cahill, L. (2017). An issue whose time has come. *Journal of neuroscience research*, 95(1-2), 12-13.

Campbell, A. (1999). Staying alive: evolution, culture and women's aggression. *Behavioral and Brain Sciences*, *22*, 203-252.

Campbell, S. M., & Collaer, M. L. (2009). Stereotype threat and gender differences in performance on a novel visuospatial task. *Psychology of Women Quarterly*, *33*(4), 437-444.

Campbell, A., & Muncer, S. (2009). Can 'risky'impulsivity explain sex differences in aggression?. *Personality and Individual Differences,* 47(5), 402-406.

Carr, K., Kendal, R. L., & Flynn, E. G. (2016). Eureka!: What Is Innovation, How Does It Develop, and Who Does It?. *Child development*.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of" playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, *1*(1), 4-17.

Charness, G., & Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, *83*(1), 50-58.

Cheng, J. T., Tracy, J. L., Foulsham, T., Kingstone, A., & Henrich, J. (2013). Two ways to the top: Evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. *Journal of Personality and Social Psychology,* 104(1), 103.

Clarke, E. & Heyes, C. (2017). The swashbuckling anthropologist: Henrich on the Secret of Our Success. *Biology and Philosophy* 32 (2):289-305 (2017)

Cole, H. W., Figler, M. H., Parente, F. J., & Peeke, H. V. (1980). The relationship between sex and aggression in convict cichlids (Cichlasoma nigrofasciatum Günther). *Behaviour*, *75*(1), 1-21.

Cooke-Simpson, A., & Voyer, D. (2007). Confidence and gender differences on the Mental Rotations Test. *Learning and Individual Differences, 17(2),* 181-186.

Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of personality and social psychology*, *37*(1), 131.

Costa Jr, P., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: robust and surprising findings. *Journal of personality and social psychology*, 81(2), 322.

Croizet, J. C., Després, G., Gauzins, M. E., Huguet, P., Leyens, J. P., & Méot, A. (2004). Stereotype threat undermines intellectual performance by triggering a disruptive mental load. *Personality and Social Psychology Bulletin*, *30*(6), 721-731.

Cross, C. P., Brown, G. R., Morgan, T. J., & Laland, K. N. (2016). Sex differences in confidence influence patterns of conformity. *British Journal of Psychology,*

Cross, C. P., Copping, L. T., & Campbell, A. (2011). Sex differences in impulsivity: a meta-analysis. *Psychological bulletin*, *137*(1), 97.

Cross, C. P., Cyrenne, D. L. M., & Brown, G. R. (2013). Sex differences in sensation-seeking: a meta-analysis. *Scientific reports*, *3*.

Dahlbom, L., Jakobsson, A., Jakobsson, N., & Kotsadam, A. (2011). Gender and overconfidence: are girls really overconfident?. *Applied Economics Letters*, *18*(4), 325-327.

Daly, M., & Wilson, M. (1988). *Homicide*. Transaction Publishers.

Davies, M., & Stone, T. (1995). Folk psychology: The theory of mind debate.

Davies, P. L., & Rose, J. D. (2000). Motor skills of typically developing adolescents: awkwardness or improvement?. *Physical & occupational therapy in pediatrics*, *20*(1), 19-42.

Dean, L. G., Hoppitt, W., Laland, K. N., & Kendal, R. L. (2011). Sex ratio affects sex-specific innovation and learning in captive ruffed lemurs (Varecia variegata and Varecia rubra). *American journal of primatology*, *73*(12), 1210-1221.

Dean, L. G., Kendal, R. L., Schapiro, S. J., Thierry, B., & Laland, K. N. (2012). Identification of the social and cognitive processes underlying human cumulative culture. *Science*, *335*(6072), 1114-1118.

Dean, L. G., Vale, G. L., Laland, K. N., Flynn, E., & Kendal, R. L. (2014). Human cumulative culture: a comparative perspective. *Biological Reviews*, *89*(2), 284-301.

Derex, M., Feron, R., Godelle, B., & Raymond, M. (2015). Social learning and the replication process: an experimental investigation. In *Proc. R. Soc. B* (Vol. 282(1808), 20150719.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, *51*(3), 629.

Dorfberger, S., Adi-Japha, E., & Karni, A. (2009). Sex differences in motor performance and motor learning in children and adolescents: an increasing male advantage in motor learning and consolidation phase gains. *Behavioural brain research*, *198*(1), 165-171.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PloS one*, 7(1), e29081.

Doyle, R. A., & Voyer, D. (2016). Stereotype manipulation effects on math and spatial test performance: A meta-analysis. *Learning and Individual Differences*, 47, 103-116.

Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics,* 56(2), 455-463.

Eagly, A. H., & Carli, L. L. (1981). Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: A meta-analysis of social influence studies. *Psychological Bulletin*, *90*(1), 1.

Eagly, A. H., Wood, W., & Fishbaugh, L. (1981). Sex differences in conformity: Surveillance by the group as a determinant of male nonconformity. *Journal of Personality and Social Psychology*, *40*(2), 384.

Eals, M., & Silverman, I. (1994). The hunter-gatherer theory of spatial sex differences: Proximate factors mediating the female advantage in recall of object arrays. *Ethology and Sociobiology*, *15*(2), 95-105.

Efferson, C., Lalive, R., Richerson, P. J., McElreath, R., & Lubell, M. (2008). Conformists and mavericks: the empirics of frequency-dependent cultural transmission. *Evolution and Human Behavior*, *29*(1), 56-64.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, 315(7109), 629-634.

Enserink, M. (2012). Final report on Stapel also blames field as a whole. *Science*, 338(6112), 1270-1271.

Eriksson, K., & Strimling, P. (2009). Biases for acquiring information individually rather than socially. *Journal of Evolutionary Psychology*, *7*(4), 309-329.

Estes, Z., & Felker, S. (2012). Confidence mediates the sex difference in mental rotation performance. *Archives of sexual behavior*, 41(3), 557-570.

Eysenck, S. B., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and individual differences*, *6*(1), 21-29.

Feldman, M. W., Aoki, K., & Kumm, J. (1996). Individual versus social learning: evolutionary analysis in a fluctuating environment. *Anthropological Science*, *104*(3), 209-231.

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological science*, *18*(10), 850-855.

Fischer, M. R. (2015). Replication–The ugly duckling of science? *GMS Z Med Ausbild*, 32, 5.

Fiske, S. (2016). *A Call to Change Science's Culture of Shaming.* Observer Magazine. from: www.psychologicalscience.org/observer

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*.

Flynn, E., & Whiten, A. (2008). Cultural transmission of tool use in young children: A diffusion chain study. *Social Development*, *17*(3), 699-718.

Flynn, E., Turner, C., & Giraldeau, L. A. (2016). Selectivity in social and asocial learning: investigating the prevalence, effect and development of young children's learning preferences. *Phil. Trans. R. Soc. B*, *371*(1690), 20150189.

Flynn, J., Slovic, P., & Mertz, C. K. (1994). Gender, race, and perception of environmental health risks. *Risk analysis*, *14*(6), 1101-1108.

Forstmeier, Wolfgang, Eric-Jan Wagenmakers, and Timothy H. Parker. "Detecting and avoiding likely false-positive findings–a practical guide." Biological Reviews (2016).

Furnham, A. (2001). Self-estimates of intelligence: Culture and gender difference in self and other estimates of both general (g) and multiple intelligences. *Personality and Individual Differences*, *31*(8), 1381-1405.

Furnham, A., von Stumm, S., Makendrayogam, A., & Chamorro-Premuzic, T. (2009). A Taxonomy of Self-Estimated Human Performance: The General Factor. *Journal of Individual Differences*, *30*(4), 188-193.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997-1016.

Gelman, A. (2016, September 21). *What has happened down here is the winds have changed*. from http://andrewgelman.com/2016

Gelman, A. (2016, Feburary 12). *Priming Effects Replicate Just Fine, Thanks.* from www.andrewgelman.com/2012

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics,* 15(3), 373-390.

Gerstenberg, F. X., Imhoff, R., & Schmitt, M. (2012). 'Women are Bad at Math, but I'm Not, am I?'Fragile Mathematical Self-concept Predicts Vulnerability to a Stereotype Threat Effect on Mathematical Performance. *European Journal of Personality*, *26*(6), 588-599.

Gilbert, D. T., King, G., Pettigrew, S. & Wilson, T. D. (2016) Comment on "Estimating the reproducibility of psychological science". *Science* **351**, 1037.

Glasser, H. M., & Smith, J. P., III. (2008). On the vague meaning of "gender" in education research: The problem, its sources, and recommendations for practice. Educational Researcher, 37, 343– 350. doi:10.3102/0013189X08323718.

Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, *77*(5), 1637-1664.

Goldberg, C. (1975). Conformity to majority type as a function of task and acceptance of sex-related stereotypes. *The Journal of Psychology*, *89*(1), 25-37.

Guadagno, R. E., & Cialdini, R. B. (2007). Persuade him by email, but see her in person: Online persuasion revisited. *Computers in Human Behavior*, *23*(2), 999-1015.

Guenther, A., Brust, V., Dersen, M., & Trillmich, F. (2014). Learning and personality types are related in cavies (Cavia aperea). *Journal of comparative psychology*, *128*(1), 74.

Greenberg, R. (2003). The role of neophobia and neophilia in the development of innovative behaviour of birds.

Grimshaw, G. M., Sitarenios, G., & Finegan, J. A. K. (1995). Mental rotation At 7 years-relations with prenatal testosterone levels and spatial play experiences. *Brain and cognition*, *29*(1), 85-100.

Gruber, T., Clay, Z., & Zuberbühler, K. (2010). A comparison of bonobo and chimpanzee tool use: evidence for a female bias in the Pan lineage. *Animal Behaviour*, *80*(6), 1023-1033.

Guadagno, R. E., & Cialdini, R. B. (2002). Online persuasion: An examination of gender differences in computer-mediated interpersonal influence. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 38.

Guadagno, R. E., & Cialdini, R. B. (2007). Persuade him by email, but see her in person: Online persuasion revisited. *Computers in Human Behavior*, *23*(2), 999-1015.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., & Zwienenberg, M. (2016). A multi-lab pre-registered replication of the ego-depletion effect. Perspectives on Psychological Science, 2.

Harris, C. R., Jenkins, M., & Glaser, D. (2006). Gender differences in risk assessment: why do women take fewer risks than men?. *Judgment and Decision Making*, *1*(1), 48.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., & Zwienenberg, M. (2015). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 2.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, *48*(1), 400-407.

Hausmann, M. (2014). Arts versus science—Academic background implicitly activates gender stereotypes on cognitive abilities with threat raising men's (but lowering women's)

Hausmann, M., Schoofs, D., Rosenthal, H. E., & Jordan, K. (2009). Interactive effects of sex hormones and gender stereotypes on cognitive sex differences—a psychobiosocial approach. *Psychoneuroendocrinology*, *34*(3), 389-401.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication monographs*, *76*(4), 408-420.

Heil, M., Jansen, P., Quaiser-Pohl, C., & Neuburger, S. (2012). Gender-specific effects of artificially induced gender beliefs in mental rotation. *Learning and Individual Differences*, *22*(3), 350-353.

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. Psychological Review, 18, 186–201.

Henrich, J. (2015). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Henrich, J., & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and human behavior*, *19*(4), 215-241.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29-29.

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *12*(3), 123-135.

Heyes, C. (2012). What's social about social learning?. *Journal of Comparative Psychology*, *126*(2), 193.

Heyes, C. (2016). Who knows? Metacognitive social learning strategies. *Trends in cognitive sciences*, *20*(3), 204-213.

Hirnstein, M., Andrews, L. C., & Hausmann, M. (2014). Gender-stereotyping and cognitive sex differences in mixed-and same-sex groups. *Archives of sexual behavior*, *43*(8), 1663-1673.

Hirnstein, M., Freund, N., & Hausmann, M. (2012). Gender stereotyping enhances verbal fluency performance in men (and women). *Zeitschrift für Psychologie*, *220*(2), 70.

Hoppitt, W., & Laland, K. N. (2013). *Social learning: an introduction to mechanisms, methods, and models*. Princeton University Press.

Horton, R. (2015). Offline: What is medicine's 5 sigma. *The Lancet*, 385(9976), 1380.

Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, *2*(3), 2053168015604648.

Humle, T., Snowdon, C. T., & Matsuzawa, T. (2009). Social influences on ant-dipping acquisition in the wild chimpanzees (Pan troglodytes verus) of Bossou, Guinea, West Africa. *Animal cognition*, *12*(1), 37-48.

Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T.D., Elliott, M.A., Ruparel, K., Hakonarson, H., Gur, R.E., Gur, R.C. & Verma, R. (2014). Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences*, 111(2), pp.823-828. Vancouver

Inzlicht, M., Aronson, J., Good, C., & McKay, L. (2006). A particular resiliency to threatening environments. *Journal of Experimental Social Psychology*, *42*(3), 323-336.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), e124.

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4(3), 245-253.

Jacobs, R. C., & Campbell, D. T. (1961). The perpetuation of an arbitrary tradition through several generations of a laboratory microculture. *The Journal of Abnormal and Social Psychology*, *62*(3), 649.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language*, 59(4), 434-446.

Jakobsson, N. (2012). Gender and confidence: are women underconfident? Applied Economics Letters, 19(11), 1057-1059.

Jakobsson, N., Levin, M., & Kotsadam, A. (2013). Gender and overconfidence: effects of context, gendered stereotypes, and peer group. *Advances in Applied Sociology*, *3*(02), 137.

Johnson, D. M. (1939). Confidence and speed in the two-category judgment. Archives of Psychology, 241, 1–52.

Johnson, D. D., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, 477(7364), 317-320.

Johnson, W., Zava, D., & McCoy, N. (2000). Overall self-confidence, self-confidence in mathematics, and sex-role stereotyping in relation to salivary free testosterone in university women. *Perceptual and motor skills,* 91(2), 391-401.

Jones, P. L., Ryan, M. J., Flores, V., & Page, R. A. (2013). When to approach novel prey cues? Social learning strategies in frog-eating bats. *Proceedings of the Royal Society of London B: Biological Sciences*, *280*(1772), 20132330.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct?. *Journal of personality and social psychology*, 83(3), 693.

Kahan, D. M., Braman, D., Gastil, J., Slovic, P., & Mertz, C. K. (2007). Culture and identity-protective cognition: Explaining the white-male effect in risk perception. *Journal of Empirical Legal Studies*, *4*(3), 465-505.

Kahneman D. A new etiquette for replication. Social Psychology. 2014; 45:310–311

Kameda, T., & Nakanishi, D. (2002). Cost–benefit analysis of social/cultural learning in a nonstationary uncertain environment: An evolutionary simulation and an experiment with human subjects. *Evolution and Human Behavior*, *23*(5), 373-393.

Kendal, R. L., Coolen, I., van Bergen, Y., & Laland, K. N. (2005). Trade-offs in the adaptive use of social and asocial learning. *Advances in the Study of Behavior*, *35*, 333-379.

Kerr, Norbert L. "HARKing: Hypothesizing after the results are known." Personality and Social Psychology Review 2, no. 3 (1998): 196-217.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., ... & Errington, T. M. (2016). Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. *PLoS Biol*, *14*(5), e1002456.

King, A. J., & Cowlishaw, G. (2007). When to use social information: the advantage of large group size in individual decision making. *Biology Letters,* 3(2), 137-139.

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences*, *17*(2), 161-173.

Koenig, A. M., & Eagly, A. H. (2005). Stereotype threat in men on a test of social sensitivity. *Sex Roles*, *52*(7-8), 489-496.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science,* 6(3), 299-312.

Laland, K. N. (2004). Social learning strategies. Animal Learning & Behavior, 32(1), 4-14.

Laland, K. N. (2017). *Darwin's Unfinished Symphony: How Culture Made the Human Mind*. Princeton University Press.

Laland, K. N., & Reader, S. M. (1999). Foraging innovation is inversely related to competitive ability in male but not in female guppies. *Behavioral Ecology*, *10*(3), 270-274.

Laland, K. N., Uller, T., Feldman, M. W., Sterelny, K., Müller, G. B., Moczek, A., ... & Odling-Smee, J. (2015, August). The extended evolutionary synthesis: its structure, assumptions and predictions. In *Proc. R. Soc. B* (Vol. 282, No. 1813, p. 20151019). The Royal Society.

Laland, K., Wray, G. A., & Hoekstra, H. E. (2014). Does evolutionary theory need a rethink?. *Nature*, *514*(7521), 161.

Langan, A. M., Shuker, D. M., Cullen, W. R., Penney, D., Preziosi, R. F., & Wheater, C. P. (2008). Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. *Assessment & Evaluation in Higher Education*, *33*(2), 179-190.

Latané, B., & Wolf, S. (1981). The social impact of majorities and minorities. *Psychological Review*, *88*(5), 438.

Lee, E. J. (2006). When and how does depersonalization increase conformity to group norms in computer-mediated communication?. *Communication Research*, *33*(6), 423-447.

Lee, M. D., & Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: comment on Trafimow (2003).

Leyens, J. P., Désert, M., Croizet, J. C., & Darcis, C. (2000). Stereotype threat: are lower status and history of stigmatization preconditions of stereotype threat? *Personality and Social Psychology Bulletin*, *26*(10), 1189-1199.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational behavior and human performance*, *20*(2), 159-183.

Lohman, D. F. (1986). The effect of speed-accuracy tradeoff on sex differences in mental rotation. *Perception & Psychophysics*, *39*(6), 427-436.

Lonsdorf, E. V., Eberly, L. E., & Pusey, A. E. (2004). Sex differences in learning in chimpanzees. *Nature*, *428*(6984), 715-716.

Mace, R., & Holden, C. J. (2005). A phylogenetic approach to cultural evolution. *Trends in ecology & evolution*, *20*(3), 116-121.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annu. Rev. Psychol.*, *58*, 593-614.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological methods*, *7*(1), 83.

Maney, D. L. (2016). Perils and pitfalls of reporting sex differences. *Phil. Trans. R. Soc. B*, *371*(1688), 20150119.

Mann, J., Sargeant, B. (2003). Like mother, like calf: the ontogeny of foraging traditions in wild Indian Ocean bottlenose dolphins (Tursiops sp.). *The biology of traditions: Models and evidence*, 236-266.

Marx, D. M., & Ko, S. J. (2012). Superstars "like" me: The effect of role model similarity on performance under threat. *European Journal of Social Psychology*,*42*(7), 807-812.

Massen, J. J., Antonides, A., Arnold, A. M. K., Bionda, T., & Koski, S. E. (2013). A behavioral view on chimpanzee personality: Exploration tendency, persistence, boldness, and tool-orientation measured with group experiments. American Journal of Primatology, 75(9), 947-958.

McElreath, R. (2016). Statistical rethinking: A Bayesian course with examples in R and Stan (Vol. 122). CRC Press.

McElreath, R., & Smaldino, P. E. (2015). Replication, communication, and the population dynamics of scientific discovery. *PloS one*, *10*(8), e0136088.

McGlone, M. S., & Aronson, J. (2006). Stereotype threat, identity salience, and spatial reasoning. *Journal of Applied Developmental Psychology*, *27*(5), 486-493.

McGrew, W. C., Tutin, C. E., & Midgett, P. S. (1975). Tool use in a group of captive chimpanzees I. Escape. *Zeitschrift fuer Tierpsychologie*, *37*(2), 145-162.

Mesoudi, A. (2009). How cultural evolutionary theory can inform social psychology and vice versa. *Psychological review*, *116*(4), 929.

Mesoudi, A. (2011). Variable cultural acquisition costs constrain cumulative cultural evolution. *PloS one*, *6*(3), e18239.

Mesoudi, A. (2016). Cultural evolution: Integrating psychology, evolution and culture. *Current Opinion in Psychology*, *7*, 17-22.

Mesoudi, A., Chang, L., Dall, S. R., & Thornton, A. (2016). The Evolution of Individual and Cultural Variation in Social Learning. Trends in ecology & evolution, 31(3), 215-225.

Mesoudi, A., Chang, L., Murray, K., & Lu, H. J. (2015). Higher frequency of social learning in China than in the West shows cultural variation in the dynamics of cultural evolution. *Proceedings of the Royal Society of London B: Biological Sciences*, *282*(1798), 20142209.

Mesoudi, A., & Lycett, S. J. (2009). Random copying, frequency-dependent copying and culture change. *Evolution and Human Behavior*, *30*(1), 41-48.

Mesoudi, A., & O'Brien, M. J. (2008). The cultural transmission of Great Basin projectile-point technology I: an experimental simulation. *American Antiquity*, 3-28.

Mesoudi, A., & Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*(1509), 3489-3501.

Mesoudi, A., Whiten, A., & Laland, K. N. (2004). Perspective: Is human cultural evolution Darwinian? Evidence reviewed from the perspective of The Origin of Species. *Evolution*, *58*(1), 1-11.

Mesoudi, A., Whiten, A., & Laland, K. N. (2006). Towards a unified science of cultural evolution. *Behavioral and Brain Sciences*, *29*(04), 329-347.

Miller, D. I., & Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in cognitive sciences*, *18*(1), 37-45.

Mishra, S., Hing, L. S. S., & Lalumière, M. L. (2015). Inequality and risk-taking. *Evolutionary Psychology,* 13(3), 1474704915596295.

Moè, A. (2009). Are males always better than females in mental rotation? Exploring a gender belief explanation. *Learning and Individual Differences*, *19*(1), 21-27.

Moè, A., & Pazzaglia, F. (2006). Following the instructions!: Effects of gender beliefs in mental rotation. *Learning and Individual Differences*, *16*(4), 369-377.

Mohr, C., Rowe, A. C., & Blanke, O. (2010). The influence of sex and empathy on putting oneself in the shoes of others. *British Journal of Psychology,* 101(2), 277-291.

Morgan, T. J. H., Rendell, L. E., Ehn, M., Hoppitt, W., & Laland, K. N. (2012). The evolutionary basis of human social learning. *Proceedings of the Royal Society of London B: Biological Sciences,* 279(1729), 653-662.

Mori, K., & Arai, M. (2010). No need to fake it: Reproduction of the Asch experiment without confederates. *International Journal of Psychology*, *45*(5), 390-397.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41), 16474-16479.

Muehlenhard, C. L., & Peterson, Z. D. (2011). Distinguishing between sex and gender: History, current conceptualizations, and implications. *Sex Roles*, *64*(11-12), 791-803.

Muthukrishna, M., Heine, S. J., Toyakawa, W., Hamamura, T., Kameda, T., & Henrich, J. (2014). Overconfidence is universal? Depends on what you mean. *Manuscript submitted for publication.*

Nelson, J. A. (2015). Are Women Really More Risk-Averse than Men? A Re-Analysis of the Literature Using Expanded Methods. *Journal of Economic Surveys*, *29*(3), 566-585.

Nguyen, H. H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology,* 93(6), 1314.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

Orsini, C. A., & Setlow, B. (2017). Sex differences in animal models of decision making. *Journal of Neuroscience Research*, *95*(1-2), 260-269.

Orsini, C. A., Willis, M. L., Gilbert, R. J., Bizon, J. L., & Setlow, B. (2016). Sex differences in a rat model of risky decision making. *Behavioral neuroscience*, *130*(1), 50.

Ortner, T. M., & Sieverding, M. (2008). Where are the gender differences? Male priming boosts spatial skills in women. *Sex roles*, *59*(3-4), 274-281.

Pajares, F., Miller, M. D., & Johnson, M. J. (1999). Gender differences in writing self-beliefs of elementary school students. *Journal of educational Psychology*, *91*(1), 50.

Pajares, F., & Valiante, G. (2001). Gender differences in writing motivation and achievement of middle school students: A function of gender orientation?. *Contemporary educational psychology*, *26*(3), 366-381.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G.,Stankov, L., et al. (2002). Individual differences in the realism of confidence judgments. *Journal of General Psychology,* 129, 257–300.

Peters, M. (2005). Sex differences and the factor of time in solving Vandenberg and Kuse mental rotation problems. *Brain and Cognition,* 57, 176–184.

Peterson, D. (2016). The baby factory difficult research objects, disciplinary standards, and the production of statistical significance. *Socius: Sociological Research for a Dynamic World*, 2, 2378023115625071.

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic bulletin & review*, *10*(1), 177-183.

Petrusic, W. M., & Baranski, J. V. (2009). Probability assessment with response times and confidence in perception and knowledge. *Acta psychologica*, *130*(2), 103-114.

Picho, K., Rodriguez, A., & Finnie, L. (2013). Exploring the moderating role of context on the mathematics performance of females under stereotype threat: A meta-analysis. *The Journal of social psychology*, *153*(3), 299-333.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184-188.

Prager, E. M. (2017). Addressing sex as a biological variable. *Journal of Neuroscience Research*, 95(1-2), 11-11.

Pulverer, B. (2015). Reproducibility blues. *The EMBO journal*, 34(22), 2721-2724.

Quaiser-Pohl, C., Geiser, C., & Lehmann, W. (2006). The relationship between computer-game preference, gender, and mental-rotation ability. *Personality and Individual differences*, *40*(3), 609-619.

Rafacz, M., & Templeton, J. J. (2003). Environmental unpredictability and the value of social information for foraging starlings. *Ethology*, *109*(12), 951-960.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the Robustness of Power Posing No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women. Psychological science, 0956797614553946.

Reader, S. M., & Laland, K. N. (Eds.). (2003). *Animal innovation* (Vol. 10). Oxford: Oxford University Press.

Réale, D., Reader, S. M., Sol, D., McDougall, P. T., & Dingemanse, N. J. (2007). Integrating animal temperament within ecology and evolution. *Biological reviews,* 82(2), 291-318.

Reilly, D., & Neumann, D. L. (2013). Gender-role differences in spatial ability: A meta-analytic review. *Sex Roles*, *68*(9-10), 521-535.

Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., ... & Laland, K. N. (2010). Why copy others? Insights from the social learning strategies tournament. *Science*, *328*(5975), 208-213.

Resnick, S. M. (1993). Sex differences in mental rotations: an effect of time limits? *Brain and Cognition, 21,* 71–79.

Rogers, A. R. (1988). Does biology constrain culture?. *American Anthropologist*, *90*(4), 819-831.

Rosander, M., & Eriksson, O. (2012). Conformity on the Internet–The role of task difficulty and gender differences. *Computers in human behavior*, *28*(5), 1587-1595.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin,* 86(3), 638.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology,* 56(5), 356-374.

Ronay, R., & Kim, D. Y. (2006). Gender differences in explicit and implicit risk attitudes: A socially facilitated phenomenon. *British Journal of Social Psychology*, *45*(2), 397-419.

Rothgerber, H., & Wolsiefer, K. (2014). A naturalistic study of stereotype threat in young female chess players. *Group Processes & Intergroup Relations*, *17*(1), 79-90.

Rubin, G. (1975). The traffic in women: Notes on the “political economy” of sex. In R. R. Reiter (Ed.), Toward an anthropology of women (pp. 157–210). New York: Monthly Review Press.

Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass*, *5*(6), 359-371.

Rydell, R. J., Rydell, M. T., & Boucher, K. L. (2010). The effect of negative performance stereotypes on learning. *Journal of personality and social psychology*, *99*(6), 883.

Rzhetsky, A., Foster, J. G., Foster, I. T., & Evans, J. A. (2015). Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, *112*(47), 14569-14574.

Sample, I. (2013). *Male and female brains wired differently, scans reveal.* The Guardian. from: www.theguardian.com

Schaefer, P. S., Williams, C. C., Goodie, A. S., & Campbell, W. K. (2004). Overconfidence and the big five. *Journal of research in Personality*, 38(5), 473-480.

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles*, 57(7-8), 509-514.

Schnall, S. (2014). Clean data: Statistical artifacts wash out replication efforts. *Social Psychology*, 45(4), 315-317

Schneeberger, N. A., & Williams, K. (2003). Why women “can't” do math: The role of cognitive load in stereotype threat research. In *18th meeting of the Society for Industrial and Organizational Psychology, Orlando, FL*.

Schroeder, J., Dugdale, H.L., Radersma, R., Hinsch, M., Buehler, D.M., Saul, J., Porter, L., Liker, A., Cauwer, I., Johnson, P.J. and Santure, A.W. (2013). Fewer invited talks by women in evolutionary biology symposia. *Journal of Evolutionary Biology*, 26(9), 2063-2069.

Sharps, M. J., Price, J. L., & Williams, J. K. (1994). Spatial Cognition and Gender Instructional and Stimulus Influences on Mental Image Rotation Performance. *Psychology of Women Quarterly*, 18(3), 413-425.

Shepard, R. N., & Metzler, J. (1971). Mental Rotation of 3-Dimensional Objects. *Science*, *171*, 3972.

Shepard, S., & Metzler, D. (1988). Mental rotation: effects of dimensionality of objects and type of task. *Journal of Experimental Psychology: Human Perception and Performance,* 14(1), 3.

Sherif, M. (1936). The psychology of social norms. Oxford, England: Harper.

Silverman, I., Choi, J., & Peters, M. (2007). The hunter-gatherer theory of sex differences in spatial abilities: Data from 40 countries. *Archives of sexual behavior*, *36*(2), 261-268.

Smaldino, P. E., & McElreath, R. (2016). The Natural Selection of Bad Science. *Royal Society Open Science,* 3(9),160384.

Smith, B. R., & Blumstein, D. T. (2008). Fitness consequences of personality: a meta-analysis. *Behavioral Ecology*, *19*(2), 448-455.

Smolla, M., Alem, S., Chittka, L., & Shultz, S. (2016). Copy-when-uncertain: bumblebees rely on social information when rewards are highly variable. *Biology Letters*, *12*(6), 20160188.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology*, *35*(1), 4-28.

Stankov, L. (2000). Complexity, metacognition, and fluid intelligence. *Intelligence*, *28*(2), 121-143.

Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence*, *25*(2), 93-109.

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology*, *100*(4), 961.

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences*, *22*(6), 747-758.

Stanovich, K. E. (1999). Who is rational? Studies of individual differences in reasoning. London: Erlbaum.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology*, *69*(5), 797.

Steinberg, J. R., Okun, M. A., & Aiken, L. S. (2012). Calculus GPA and Math Identification as Moderators of Stereotype Threat in Highly Persistent Women. *Basic and Applied Social Psychology*, *34*(6), 534-543.

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? Review of General Psychology, 16(1), 93

Sutter, M., & Glätzle-Rützler, D. (2014). Gender differences in the willingness to compete emerge early in life and persist. *Management Science*.

Syzmanowicz, A., & Furnham, A. (2011). Gender differences in self-estimates of general, mathematical, spatial and verbal intelligence: Four meta analyses. Learning and Individual Differences, 21(5), 493-504.

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual review of psychology*, *33*(1), 1-39.

Tebbich, S., Stankewitz, S., & Teschke, I. (2012). The relationship between foraging, learning abilities and neophobia in two species of Darwin's finches. *Ethology*, *118*(2), 135-146.

Tellegen, A. (1982). *Brief manual for the Differential Personality Questionnaire.* Unpublished manuscript, Department of Psychology, University of Minnesota Press, Minneapolis

Toelch, U., Bruce, M. J., Meeus, M. T., & Reader, S. M. (2011). Social performance cues induce behavioral flexibility in humans. Frontiers in psychology, 2, 160.

Toelch, U., Bruce, M. J., Newson, L., Richerson, P. J., & Reader, S. M. (2014). Individual consistency and flexibility in human social information use. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1776), 20132864.

Torgrimson, B. N., & Minson, C. T. (2005). Sex and gender: what is the difference? *Journal of Applied Physiology 787-787*

Thornton, A., & Samson, J. (2012). Innovative problem solving in wild meerkats. *Animal Behaviour*, *83*(6), 1459-1468.

Trafimow, D, Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37(1),1-2.

Trivers, R. (1972). Parental investment and sexual selection. *Sexual Selection & the Descent of Man, Aldine de Gruyter, New York*, 136-179.

Trix, F., & Psenka, C. (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse & Society,* 14(2), 191-220.

Unger, R. K. (1979). Toward a redefinition of sex and gender. American Psychologist, 34, 1085–1094. doi:10.1037/0003-066X.34.11.1085.

van Bergen, Y., Coolen, I., & Laland, K. N. (2004). Nine-spined sticklebacks exploit the most reliable source when public and private information conflict. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, *271*(1542), 957-962.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: applications to developmental research. *Child development*, *85*(3), 842-860.

van de Waal, E., Borgeaud, C., & Whiten, A. (2013). Potent social learning and conformity shape a wild primate's foraging decisions. *Science*, *340*(6131), 483-485.

van Leeuwen, E. J., Acerbi, A., Kendal, R. L., Tennie, C., & Haun, D. B. (2016). A reappreciation of 'conformity'. *Animal Behaviour*, *122*, e5-e10.

van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2-12.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and motor skills.*

Vederhus, L., & Krekling, S. (1996). Sex differences in visual spatial ability in 9-year-old children. *Intelligence*, *23*(1), 33-43.

Verbeek, M. E., Drent, P. J., & Wiepkema, P. R. (1994). Consistent individual differences in early exploratory behaviour of male great tits. *Animal Behaviour*, *48*(5), 1113-1121.

Von Stumm, S., Chamorro-Premuzic, T., & Furnham, A. (2009). Decomposing self-estimates of intelligence: Structure and sex differences across 12 nations. *British Journal of Psychology*, 100(2), 429-442.

Voyer, D., & Hou, J. (2006). Type of items and the magnitude of gender differences on the Mental Rotations Test. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *60*(2), 91.

Voyer, D., Rodgers, M. A., & McCormick, P. A. (2004). Timing conditions and the magnitude of gender differences on the Mental Rotations Test. Memory & Cognition, 32, 72–82.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychological bulletin*, *117*(2), 250.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on psychological science*, *4*(3), 274-290.

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology*, *39*(5), 456-467.

Walton, G. M., & Spencer, S. J. (2009). Latent ability grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, *20*(9), 1132-1139.

Wang, X. T. (2002). Risk as reproductive variance. *Evolution and Human Behavior*, *23*(1), 35-57.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*.

Wheeling, K. (2015) *The Brains of men and women aren't really that different, study finds.* Science. from: www.sciencemag.org. DOI: 10.1126/science.aad7499

Wilson, D. S., Clark, A. B., Coleman, K., & Dearstyne, T. (1994). Shyness and boldness in humans and other animals. *Trends in Ecology & Evolution*, *9*(11), 442-446.

Wisdom, T. N., & Goldstone, R. L. (2010, March). Social learning and cumulative innovations in a networked group. *In International Conference on Social Computing, Behavioral Modeling, and Prediction* (pp. 32-41). Springer Berlin Heidelberg.

Wolf, M., Van Doorn, G. S., Leimar, O., & Weissing, F. J. (2007). Life-history trade-offs favour the evolution of animal personalities. *Nature*, *447*(7144), 581-584.

Wout, D., Danso, H., Jackson, J., & Spencer, S. (2008). The many faces of stereotype threat: Group-and self-threat. *Journal of Experimental Social Psychology*, *44*(3), 792-799.

Wraga, M., Duncan, L., Jacobs, E. C., Helt, M., & Church, J. (2006). Stereotype susceptibility narrows the gender gap in imagined self-rotation performance. *Psychonomic Bulletin & Review*, *13*(5), 813-819.

Wraga, M., Helt, M., Jacobs, E., & Sullivan, K. (2006). Neural basis of stereotype-induced shifts in women's mental rotation performance. *Social cognitive and affective neuroscience*.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit formation. Journal of Comparative Neurology and Psychology, 18, 458–482.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310-1321.

Yim, O. S., Zhang, X., Shalev, I., Monakhov, M., Zhong, S., Hsu, M., ... & Ebstein, R. P. (2016). Delay discounting, genetic sensitivity, and leukocyte telomere length. *Proceedings of the National Academy of Sciences,* 113(10), 2780-2785.

Yoder, J. D. (2003). Women and gender: Transforming psychology (2nd ed.). Upper Saddle River: Pearson / Prentice Hall.

Zefferman, M. R. (2016). Mothers teach daughters because daughters teach granddaughters: the evolution of sex-biased transmission. *Behavioral Ecology*.

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of consumer research*, *37*(2), 197-206.

Appendix

Ethics letter of approval for experiment in Chapter 2:

University of St Andrews

University Teaching and Research Ethics Committee
Sub-committee

7 May 2015

Ethics Reference No: *Please quote this ref on all correspondence*	PS11481
Project Title:	"Build the best spaceship!" A short computer-based problem solving task
Researcher's Name:	Charlotte Brand
Supervisors:	Dr Gillian Brown and Dr Catharine Cross

Thank you for submitting your application which was considered at the Psychology & Neuroscience School Ethics Committee meeting on the 24th March 2015. The following documents were reviewed:

1.	Ethical Application Form	05/05/2015
2.	Advertisement	05/05/2015
3.	Participant Information Sheet	05/05/2015
4.	Consent Form	05/05/2015
5.	Debriefing Form	05/05/2015
6.	Experiment Script	05/05/2015
7.	Example of Participant View	05/05/2015
8.	Questionnaire	05/05/2015
9.	Data Management Plan	05/05/2015

The University Teaching and Research Ethics Committee (UTREC) approves this study from an ethical point of view. Please note that where approval is given by a School Ethics Committee that committee is part of UTREC and is delegated to act for UTREC.

Approval is given for three years. Projects, which have not commenced within two years of original approval, must be re-submitted to your School Ethics Committee.

You must inform your School Ethics Committee when the research has been completed. If you are unable to complete your research within the 3 three year validation period, you will be required to write to your School Ethics Committee and to UTREC (where approval was given by UTREC) to request an extension or you will need to re-apply.

Any serious adverse events or significant change which occurs in connection with this study and/or which may alter its ethical consideration, must be reported immediately to the School Ethics Committee, and an Ethical Amendment Form submitted where appropriate.

Approval is given on the understanding that the 'Guidelines for Ethical Research Practice' https://www.st-andrews.ac.uk/utrec/guidelines/ are adhered to.

Yours sincerely

pp.

Convenor of the School Ethics Committee

Ccs Dr Gillian Brown
Dr Catharine Cross
School Ethics Committee

School of Psychology & Neuroscience, St Mary's Quad, South Street, St Andrews, Fife KY16 9JP
Email: psyethics@st-andrews.ac.uk Tel: 01334 462071

Ethical letter of approval for experiments in Chapter 3:

University of St Andrews
from first to foremost

600 YEARS
1413 – 2013

Project Title	'Spot the difference' A computer based perception task using 3D human figures
Researcher's Name	Charlotte Brand
Supervisors	Dr Gillian Brown and Dr Kate Cross
Department/Unit	School of Psychology & Neuroscience
Ethical Approval Code (Approval allocated to Original Application)	PS10561
Original Application Approval Date	13 November 2013
Amendment Application Approval	18 June 2014

Ethical Amendment Approval

Thank you for submitting your amendment application which was considered at the Psychology & Neuroscience School Ethics Committee meeting on the 18th June 2014. The following documents were reviewed:

1. Ethical Amendment Application Form 18/06/2014
2. Advertisement 18/06/2014
3. Participant Information Sheet 18/06/2014
4. Consent Form 18/06/2014
5. Debriefing Form 18/06/2014
6. Data Management Plan 18/06/2014

The University Teaching and Research Ethics Committee (UTREC) approves this study from an ethical point of view. Please note that where approval is given by a School Ethics Committee that committee is part of UTREC and is delegated to act for UTREC.

Approval is given for three years from the original application only. Ethical Amendments do not extend this period but give permission to an amendment to the original approval research proposal only. If you are unable to complete your research within the original 3 three year validation period, you will be required to write to your School Ethics Committee and to UTREC (where approval was given by UTREC) to request an extension or you will need to re-apply. You must inform your School Ethics Committee when the research has been completed.

Any serious adverse events or significant change which occurs in connection with this study and/or which may alter its ethical consideration, must be reported immediately to the School Ethics Committee, and an Ethical Amendment Form submitted where appropriate.

Approval is given on the understanding that the 'Guidelines for Ethical Research Practice' (http://www.st-andrews.ac.uk/media/UTRECguidelines%20Feb%2008.pdf) are adhered to.

Yours sincerely

Convenor of the School Ethics Committee

Ccs School Ethics Committee
Dr Gillian Brown (Supervisor)
Dr Kate Cross (Supervisor)

School of Psychology & Neuroscience, St Mary's Quad, South Street, St Andrews, Fife KY16 9JP
Email: psyethics@st-andrews.ac.uk Tel: 01334 462071

Example of trace plots and posterior distributions for Chapter 3: